DOCUMENT RESUME

ED 142 563 95 TM 006 175

AUTHOR	Law, Alexander I.; Bronson, William H.
TITLE	Program Evaluator's Guide.
INSTITUTION	California State Dept. of Education, Sacramento.; Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	Mar 77
NOTE	335p.; For related document, see TM 006 407
AVAILABLE FROM	Educational Testing Service, Evaluation Improvement Program, Princeton, New Jersey 08540 (1-24 copies, $12.00 ea., 25 or more, $10.00 ea.)
EDRS PRICE	MF-$0.83 HC-$18.07 Plus Postage.
DESCRIPTORS	Data Collection; Educational Objectives; *Evaluation Methods; *Evaluators; *Guides; Information Utilization; Inservice Education; Instructional Materials; *Measurement Techniques; *Planning; *Program Evaluation; Sampling; Statistical Analysis; Test Construction; Test Interpretation; Test Results; Test Selection
IDENTIFIERS	California; *California Evaluation Improvement Project

ABSTRACT

This guide presents detailed information concerning the purposes and process of program evaluation, the role of the evaluator, and the development of an evaluation plan or design. Instruction is provided in selecting or developing assessment instruments, collecting and analyzing data, reporting evaluation results and applying the findings. The manual, which includes learning exercises, was developed under the California Evaluation Improvement Project as a study guide for use in inservice training workshops for program evaluators, teachers, principals, curriculum specialists and other individuals responsible for school programs and those who aid educational administrators ascertain program effectiveness. (EVH)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED142563

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

PROGRAM EVALUATOR'S GUIDE

TM006 175

The Evaluation Improvement Program

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Educational Testing Service
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

CALIFORNIA STATE DEPARTMENT OF EDUCATION
WILSON RILES, SUPERINTENDENT

PROGRAM EVALUATOR'S GUIDE

a manual developed as part of the California Evaluation Improvement Project under the leadership of the California State Department of Education, Wilson Riles, Superintendent of Public Instruction and Director of Education

Alexander I. Law, Chief
Office of Program Evaluation and Research

William H. Bronson, State Director
California Evaluation Improvement Project

The Evaluation Improvement Program

The activity which is the subject of this report was supported in whole or in part by the U.S. Office of Education, Department of Health, Education and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the U.S. Office of Education, and no official endorsement by the U.S. Office of Education should be inferred.

Copyright © 1977 by the California State Department of Education. All rights reserved.

First Edition, March 1977

Published by Educational Testing Service, Princeton, New Jersey

Educational Testing Service, a nonprofit organization, is

an Equal Opportunity Employer

FOREWORD

Evaluation of school programs is becoming more of a necessity for survival than a luxury enjoyed only by affluent districts. As financial resources diminish, decisions on how to allocate the available funds must be made. While basic educational research provides much valuable information, that information is usually not the kind on which day-to-day decisions about specific educational programs are based. Program evaluation, as perceived by the California Evaluation Improvement Project, is a means by which useful information is collected and analyzed by a local educational agency for its own use.

While most educators have had courses in testing and measurement and some contact with educational research, there has been little in their training to prepare them for conducting a systematic evaluation of a local school or classroom program. Of course, evaluation has been going on for many years, but it has most frequently been at the intuitive level, with little consistency and little impact on the total educational program.

California's response to this problem has been to develop a training program in basic evaluation concepts and skills, which is directed to the classroom teacher, the principal, the curriculum director, or program manager who wants to evaluate a local program to assist in local decision making.

One of the strengths of this training program is that it was developed and field tested throughout California by a group of educators whose backgrounds were primarily in the areas of program planning, curriculum, administration, and supervision. Evaluation specialists were used extensively as consultants as the workshop training program was developed, but the emphasis has been kept on how evaluation information could help in answering questions raised by the developers, whose orientation was basically that of program managers.

There is no magic formula to solve the problems involved in educating the youth of America, but I hope that this training program in basic evaluation concepts and skills will be useful to local schools and districts as they work toward improvement of the educational process.

Wilson Riles

WILSON RILES
Superintendent of Public
Instruction

PREFACE

Program evaluation, through which a school or district evaluates its own program for its own purposes, is different from educational research. It is also different from a state testing program and from gathering information required by the state about achievement levels in specially funded programs. Program evaluation at the school or district level should be something the school or district does for itself for its own purposes, rather than something an outsider does for it.

Program evaluation should be an intergral part of the program-planning process. Provisions should be built into each program to collect information that will indicate progress towards the program's objectives, the degree of implementation of the plan, and other information required to make rational decisions about the program.

Program evaluation is of little value unless some use is made of its results. A part of the evaluation process includes identifying potential audiences for the evaluation report and finding out what kinds of information would be useful to them. Providing useful, timely information to people who can use it is one of the best ways of ensuring that the evaluation reports will be used.

These concepts are basic to the workshop materials that have been developed by the California Evaluation Improvement Project. The materials were designed to be as practical as possible for the educational practitioner, and it is our hope that the reader will find these concepts useful and will be able to apply them to future planning as well as to programs that are currently in operation.

William H Bronson

Alexander I. Law
Chief, Office of Program
Evaluation and Research

William H. Bronson
State Director, California
Evaluation Improvement Project

INTRODUCTION TO THE EVALUATION IMPROVEMENT PROGRAM

Educational Testing Service (ETS) is pleased to have been selected by the California Department of Education as publisher, under an exclusive license, of the California Evaluation Improvement Project (EIP) materials. These constitute a course of instruction for individuals responsible for school programs and for those who help educational administrators ascertain program effectiveness. At the time of initial publication, spring 1977, the materials in EIP consist of the following:

- Program Evaluator's Guide. The Guide is a basic manual which provides in considerable detail background knowledge on the steps involved in planning and carrying out a program evaluation. It is designed as a study guide and learning tool for use in inservice training workshops for program evaluators.

- Workbook on Program Evaluation. The Workbook has two purposes. It can be put to use as a learning and instructional aid while one masters the procedures, techniques, and methods of program evaluation. Used this way, it helps the practitioner summarize and put into practice the subject matter presented in the Program Evaluator's Guide. It is best used, however, as a working notebook which the trained program evaluator can use for recording his or her plans as they are made and for making notes on program and program evaluation activities and events during the course of the program year. Used in this way, it helps the program evaluator keep complete records of the important information related to the program evaluation. It will probably be most useful when an interim or end-of-year program evaluation report has to be prepared, for much of the information needed at those critical times will already have been made a matter of record in the Workbook.

- Evaluation Trainer's Guide. This volume is a companion to the Program Evaluator's Guide. It supplies background and supporting materials for use by instructors conducting program evaluation workshops. Graphic art is provided for visual aids in support of a variety of subjects.

Publication early in 1977 of the first of a continuing offering of EIP materials is consistent with ETS's long-term commitment to help advance the art of program evaluation in the elementary and secondary schools. The EIP materials are expected to go through a number of printings under the ETS imprimatur. Each successive printing will be a revised edition. Here we ask the help and cooperation of the readership.

As you, as a program evaluation practitioner, identify parts of any of these three works that could benefit from refinement and further development, or as you think of experiences that would serve to illustrate points made in any of the subject treatments, we hope that you will share your thoughts with us.

We would like to see the Evaluation Improvement Program subjected to its own program evaluation by those who use its materials. We would hope the evaluation will be formative, not summative in nature, for it is our intention to cycle evaluative comment on each edition into significant improvements in later ones. Present plans call for publication of the second edition, our first revision, late in 1977, and constructively critical comments from practitioners can be turned into refinements in print in very short order. Join with us to make the EIP materials, initially well developed by the California Evaluation Improvement Project, even better as time goes on. The California Department of Education and Educational Testing Service have joined in the common goal of making the EIP materials as practical and useful as they can be made to be.

Jack R. Childress

Jack R. Childress
Vice-President
Educational Testing Service

Wesley W. Walton

Wesley W. Walton
Program Director
Educational Testing Service

CONSULTANTS TO THE PROJECT

The following consultants worked closely with the developers of these materials and offered excellent suggestions which have been incorporated into the text:

Marvin C. Alkin, Ed.D., Director
Center for the Study of Evaluation
University of California at Los Angeles

Merlynn Bergen, Ph.D., Teaching and Research Assistant
School of Education, Educational Psychology
Stanford University, Stanford

Preston T. Bishop, Ed.D., Consultant
Division of Program Evaluation, Research and Pupil Services
Office of the Los Angeles County Superintendent of Schools

Antonio DePorcel, Ph.D., Senior Research Scientist
Behavioral Science and Technology Program
Ameican Institute for Research, Palo Alto

Annalee Elman, Ph.D., Research Assistant
School of Education, Educational Psychology
Stanford University, Stanford

J. Richard Harsh, M.A., Director
Los Angeles Office
Educational Testing Service

Roger A. Kaufman, Ph.D., Professor
United States International University

Larry E. Orcutt, Ph.D., Independent Consultant
L.E. Orcutt Associates, Incorporated

Dale M. Russell, Ed.D., Consultant
Division of Program Evaluation, Research and Pupil Services

R. Garry Shirts, Ph.D., Consultant
Simile II, Corporation

Les E. Shuck, Ed.D., Assistant Superintendent
Research and Development
Newport-Mesa Unified School District

Daniel L. Stufflebeam, Ph.D., Professor
Western Michigan University

Arlene B. Tenenbaum, Ph.D., Consultant
Educational Research Division
Xerox Corporation, Palo Alto

ADVISORY COMMITTEE, DIRECTORS AND STAFF

The following educators served as members of the California Evaluation Improvement Project State Advisory Committee and provided invaluable direction and support to the project:

Advisory Committee

- Robert W. Babcock, Ed.D., Director, Evaluation Improvement Center, Southern Section, Office of the Los Angeles County Superintendent of Schools
- Raymond M. Langley, M.A., Assistant Superintendent, San Luis Obispo County Office of Education
- Alexander I. Law, Ph.D., Chief, Office of Program Evaluation and Research, California State Department of Education
- Floyd I. Marchus, Ed.D., Superintendent, Contra Costa County Office of Education
- Donald A. MacLean, Ph.D., Assistant Superintendent, Orange County Office of Education
- Oliver "Bud" Neely, M.A., Assistant Superintendent, Shasta County Office of Education
- Nelson C. Price, Ed.D., Director, Evaluation Improvment Center, Northern Section, San Mateo County Office of Education
- William J. Zachmeier, Ph.D., Assistant Superintendent, Santa Cruz County Office of Education
- Donald C. Ziehl, Ed.D., Superintendent, La Canada Unified School District

State Department of Education

- Wliiam H. Bronson, M.A., State Project Director
- Carolyn M. Fowle, Ed.D., Consultant

Development Centers

- Robert W. Babcock, Ed.D., Director, Los Angeles County Office of Education
- John Plakos, M.S.
- Marie E. Plakos, Ed.D.

- Nelson C. Price, Ed.D., Director, San Mateo County Office of Education

- Patricia Evans, M.A.
- Carmen J. Finley, Ph.D.
- Arlen L. Kennedy, M.A.
- Alice W. Rotzel, Ph.D.

Satellite Centers

- Dean M Dennett, M.A., Director, Shasta County Office of Education

- Aniello L. Malvetti, M.A., Director, Sacramento County Office of Education

- Marti V. Halter, M.A.

- Phyllis L. McKinney, M.A., Director, Orange County Office of Education

- John Smith, M.A.
- Jeffrey A. Wells, M.A.

- Glen N. Pierson, Ph.D., Director, San Diego County Office of Education

- Rodney E. Phillis, Ph.D.

- Thomas J. Riley, Ph.D., Director, Fresno County Office of Education

- Jack M. Thompson, Ed.D., Director, Sonoma County Office of Education

- Vincent "Vic" Abata, M.A.
- Gregory A. Malone, M.A.

ACKNOWLEDGEMENTS

The initial materials were developed by the following EIP staff members:

- Robert W. Babcock, Ed.D., Director, Los Angeles Evaluation Improvement Project Development Center
- Patricia Evans, M.A., Consultant, San Mateo Evaluation Improvement Project Development Center
- Carmen J. Finley, Ph.D., Consultant, San Mateo Evaluation Improvement Project Development Center
- Arlen L. Kennedy, M.A., Consultant, San Mateo Evaluation Improvement Project Development Center
- John Plakos, M.S., Consultant, Los Angeles Evaluation Improvement Project Development Center
- Marie E. Plakos, Ed.D., Consultant, Los Angeles Evaluation Improvement Project Development Center
- Nelson C. Price, Ed.D., Director, San Mateo Evaluation Improvement Project Development Center
- Alice W. Rotzel, Ph.D., Consultant, San Mateo Evaluation Improvement Project Development Center

At the end of the first and second years of development, revisions were made by Carmen J. Finley, Carolyn M. Fowle, and William H. Bronson. Refinements were based upon extensive interviews with members of the EIP staff throughout the state who had been using the EIP materials in the conduct of workshops.

Development activities were coordinated by William H. Bronson, M.A., EIP Project Director and Carolyn M. Fowle, Ed.D., Project Consultant.

Prepublication revisions were made by Wesley W. Walton, Ed.D., Director of the Evaluation Improvement Program at ETS. Nathaniel H. Hartshorne and Estelle Bartels served as ETS editors. Joan Westoff and Terry Birch provided covers and art supervision. Marissa G. Burch and Cathy E. Snyder served as text-processing machine operators.

INFORMATION ABOUT EIP MATERIALS AND WORKSHOPS

Information about ordering Evaluation Improvement Program materials, about Evaluation Improvement Program workshops that use these materials, or about making arrangements for specially scheduled EIP workshops for local, regional, or state inservice training programs may be obtained by writing or telephoning the Evaluation Improvement Program at Educational Testing Service, Room P-069, Princeton, NJ 08540, (609) 921-9000 or at any of its regional offices listed below.

REGIONAL OFFICES OF EDUCATIONAL TESTING SERVICE

3445 Peachtree Road, NE
Suite 1040
Atlanta, Georgia 30326
(404) 262-7634

3724 Jefferson, Suite 100
Austin, Texas 78731
(512) 452-8817

1947 Center Street
Berkeley, California 94704
(415) 849-0950

2200 Merton Avenue
Room 216
Los Angeles, California 90041
(213) 254-5236

960 Grove Street
Evanston, Illinois 60201
(312) 869-7700

GPO Box 1271
San Juan, Puerto Rico 00936
(809) 763-3636, 3640, or 3760

One Dupont Circle
Suite 310
Washington, D.C. 20036
(202) 296-5930

2 Sun Life Executive Park
100 Worcester Road
Wellesley Hills, Massachusetts 02181
(617) 235-8861 or 8860

CONTENTS

Section A: DETERMINE THE EVALUATION PURPOSES AND REQUIREMENTS

Section B: DEVELOP AN EVALUATION PLAN

Section C: DETERMINE THE EVALUATION DESIGN AND DO THE SAMPLING

Section D: SELECT OR DEVELOP ASSESSMENT INSTRUMENTS

Section E: COLLECT THE DATA

Section F: ANALYZE EVALUATION DATA

Section G: REPORT EVALUATION RESULTS

Section H: APPLY EVALUATION FINDINGS

Section I: SELECTED BIBLIOGRAPHY

Section J: APPENDICES

PROGRAM EVALUATOR'S GUIDE

Section A

DETERMINE THE EVALUATION PURPOSES AND REQUIREMENTS

The Evaluation Improvement Program

PRECIS

The success of a program and of its evaluation depends to a great extent upon how clearly the evaluator understands at the start what things should be like at the end. If the schools' decision makers are to have confidence in an evaluator's answers to policy questions about program effectiveness, costs, and continuance, a number of questions must be asked at the outset and their answers clarified through the evaluation process. It is critical, then, to decide early in the evaluation process what purposes the program evaluation is expected to serve and who will be involved in defining them.

In many programs that are continued from year to year, such early planning consists of surveying the outcomes of previous years' activities and determining status and needs in the areas served by the programs. The evaluator would then ascertain what goals and objectives have been set for the program and what these mean in terms of program evaluation. The link between program planning and evaluation planning is in the formulation of program objectives into terms that are measureable and with respect to which adequate measurement information can be collected and analyzed to satisfy end-of-year evaluation requirements.

CONTENTS

Page

1. PURPOSES OF PROGRAM EVALUATION A-1
    Communicating with the Public A-1
    Providing Information to Decision Makers A-3
    Improving Existing Programs A-4
    Providing Additional Satisfaction to Participants A-4

2. OVERVIEW OF THE EVALUATION PROCESS A-5
    Definition of Program Evaluation A-5
    Types of Evaluation Data A-6
    Evaluation as an Ongoing Process A-9

3. ROLE OF THE PROGRAM EVALUATOR A-11
    External Evaluator A-11
    Internal Evaluator A-12

4. INITIAL STEPS IN EVALUATION PLANNING A-13
    Find Out What the Program Evaluation Is to Accomplish A-13
    Review Needs Assessment, Program Goals and Objectives A-13
    Separate Objectives Statements from Goals Statements A-15
    Determine That Six Components of Performance Objectives
      Are Present A-15
    Summary of Evaluation Planning Stages A-16

5. REQUIREMENTS OF PROGRAM EVALUATION A-16
    Key Questions A-16
    End-of-Year Evaluation A-20
    Interim Evaluation A-20
    Identifying Resources and Constraints A-20

SUMMARY A-22
CHECKLIST OF STEPS IN DETERMINING PURPOSES AND REQUIREMENTS . . . A-23
LEARNING EXERCISE 1: TYPES OF EVALUATION DATA A-25
LEARNING EXERCISE 2: IDENTIFICATION OF MEASURABLE OBJECTIVES . . . A-27
LEARNING EXERCISE 3: SELECTING APPROPRIATE OBJECTIVES A-29
LEARNING EXERCISE 4: MATCHING NEEDS STATEMENTS TO PROGRAM
  OBJECTIVES AND PROGRAM ACTIVITIES A-39

1. PURPOSES OF PROGRAM EVALUATION

In recent years, the pressure on public schools to evaluate and publicize the results of their educational programs has markedly increased. Response to this pressure has ranged from enthusiastic compliance to delay and avoidance. Frequently, evaluation has been envisioned as producing more risks than gains. Indeed, educators have asked: Is a more thorough and improved evaluation worth the effort?

Evaluation means different things to different people. Perceptions may be limited to individual activities such as grading students, rating teachers, examining test scores, and/or judging the effectiveness of an educational activity.

The primary emphasis of the Evaluation Improvement Program is on the evaluation of educational programs. Programs, in this context, are defined as a combination of content, personnel, activities, and resources organized so as to attain specified goals and objectives. A program may be specific to an age or grade level, a subject-matter discipline, or a type of service.

Program evaluation can serve different purposes. Four major ones, which will be discussed in this section, are:

1. Communicating with the public
2. Providing information to decision makers
3. Improving an existing program
4. Providing additional satisfaction to participants

The reader may perceive additional purposes for program evaluation as he applies the concept to his own work setting.

Communicating with the Public

Schools play to a number of audiences each of which makes evaluative judgments. These judgments usually are based on limited or partial information. Frequently, a community uses superficial newspaper and/or television reports as the basis for evaluating the effectiveness of school programs. Note the following report of reading scores as excerpted from an article published by the Los Angeles Times of December 3, 1974:

READING SCORES -- GRADE 6

California State Testing Program

District	'70-'71 Median %ile	'71-'72 Median %ile	'72-'73 Median %ile	'73-'74 Median %ile
A	98	89	14	6
B	38	19	10	7
C	97	99	97	96

Note the median percentile scores of three individual districts. Most likely the diminished reading scores as reported in Districts A and B in 1973-74 caused considerable discontent with the schools on the part of the citizens of these communities.

The public also receives information about school programs from students in the family and from other informal settings. This information may or may not be biased; however, judgments are nevertheless made based on information gleaned from such sources.

In summary, the public frequently derives its opinion of the efficacy of the educational system through partial, or at times, biased information. These judgments affect the extent of financial support for schools, the degree of freedom of instruction and the self-esteem of educators. As a consequence, the opportunities available to learners may be positively or negatively influenced.

It is, therefore, beneficial to educators, to the schools and their programs, to supply the public with comprehensive information, the best that can be pulled together. Reports to the public should be based on a full range of program objectives and should show the extent to which the objectives were realized. When this is accomplished, the public will be able to make more informed judgments about the effectiveness of school programs and what is needed to gain support for them.

One must remember that within the public there are a number of audiences, and each has unique needs for information. You should identify these various audiences and ascertain the questions they may raise about current educational programs. The audiences and their questions each need to be addessed in the program evaluation. Within the general public, one might identify (1) parents, (2) teachers, (3) students, (4) the business community, (5) the industrial community, (6) the professional community, and (7) the retirement community as somewhat separate audiences.

Providing Information to Decision Makers

Judgments made by school personnel are often critical and apt to have an immediate impact. Program evaluation, then, can be helpful in making ongoing decisions. The information it produces may be applicable through all phases of educational management ranging from assessment needs through program planning and implementation to the adjustment of objectives before repeating a program.

Educators often approach educational planning with nothing more than an intuitive sense of needs. They may proceed without validating these needs in the local setting. Likewise, many educators will have programs, objectives, and plans in mind without establishing their appropriateness for filling the needs which have been identified. To increase program effectiveness, educators need to ascertain what needs exist and determine what programs will best meet those needs.

Planning for program evaluation is an integral part of planning that program. Such preparation can serve to assure a continuing focus on the most important objectives and steady progress towards their achievement. Decisions with respect to a program and its parts to adopt, modify, expand, or discontinue are made throughout the stages of its development. If useful information is not available, arbitrary decisions will be made. With evaluation information, the quality of decisions and acceptance of changes by those involved will be improved. Systematic evaluation provides a sound basis for the decisions that are reached.

Decision makers, of course, are to be found at various levels within the school. The teacher is a decision maker within the classroom, the principal within the school, and the superintendent and board of education within the district. One must consider needs at each of these levels of decision making when gathering the needed information and developing the evaluation plan.

Improving Existing Programs

An effective program evaluation system can help ongoing programs operate more effectively by providing feedback to staff about what is happening. Frequently, the instruments used to assess program results also can be used to diagnose individual instructional needs. Lacking this information, the teacher's solution may tend toward the same instruction for everyone. With it, the teacher can individualize instruction to meet each student's needs. Relevant program evaluation information may make possible a greater degree of individualization of instruction and also more effective groupings of students for instructional purposes.

Educational programs evolve and change over time as students and their needs also evolve and change. Information about the effect of different aspects of a program on students may enable the staff to identify the factors which may need modification as the program proceeds. In a school system, there are persons at different levels who have access to different resources and who will take different actions in their attempts to improve ongoing programs. The teacher may adjust instructional methods; the principal may assign new personnel and/or resource materials; the board of education and superintendent may grant additional financial support. Each makes unique contributions and therefore has unique needs for evaluation information.

Providing Additional Satisfaction to Participants

Program evaluation differs from individual evaluation in that it measures objectives which apply to groups of persons, perhaps by grade level, academic department, or an entire school or a group of schools. Therefore, evaluation of programs can be conducted in a context of mutual help rather than posing individual threats, as sometimes occurs in teacher evaluation, or imposing too much testing, as sometimes occurs in student evaluation. Assessment of common objectives usually generates a sense of unity and growth. Program evaluation offers maximum benefits and minimum burdens for all in the schools.

Program evaluation may be designed to give useful information to students as well as to program managers. Especially if a student's progress as shown in a program evaluation is compared to his own previous performance, he is likely to see progress and feel positive about his growth. The beneficiaries in such a situation clearly are the students and the instructional staff. Questions they would like answered are important and most assuredly should be built in as part of the evaluation plan.

2. OVERVIEW OF THE EVALUATION PROCESS

Definition of Program Evaluation

Program evaluation is defined here as the process of determining the value or effectiveness of an activity for the purpose of decision making. The key words in this definition are (1) value, (2) effectiveness, and (3) decision making.

1. Value. When a program evaluation takes place, the decision maker is concerned with determining the net value of something, its costs in relation to its benefits. Both costs and benefits have to be measured in terms of human factors and dollars.
2. Effectiveness. The decision maker needs to know to what extent a particular program was effective in meeting identified needs or objectives. Measures of effectiveness tell the decision maker what difference the program has made.

3. *Decision making*. A person with program responsibilites needs information on value and effectiveness which is useful in deciding what to do next: to continue, modify, or drop a program. The purpose of program evaluation is to improve the quality of the program decisions reached.

The evaluation process divides itself into three major phases: (1) Planning, (2) Conducting, and (3) Using. Each of the phases has distinct components. See the chart below:

THE EVALUATION PROCESS

PLAN	CONDUCT	USE
• Determine Evaluation Purpose and Objectives	• Collect Data	• Report Results
• Develop the General Evaluation Plan	• Analyze Data	• Apply Evaluation Findings
• Determine the Specific Evaluation Design		
• Obtain Assessment Tools		

This *Guide* is based on these eight components, with a section devoted to each. This section focuses on the first step, "Determine the Evaluation Purpose and Objectives."

Types of Evaluation Data

There are two basic types of program evaluation with which educators are concerned. They are formative and summative.

- Formative evaluation takes place during the development of a program or instructional unit. It is concerned with fine tuning the implementation processes and measuring learner progress as the program moves toward the attainment of specified objectives. Thus, formative evaluation provides the decision maker with information during the course of program development and execution for possible midcourse corrections to help assure that the program objectives are eventually met in an effective and economical fashion.

- Summative evaluation takes place at the end of a program or an instructional unit. This type of evaluation is concerned with measuring levels of learner achievement and the success (or failure) of operational procedures.

The two types of evaluation along with three kinds of evaluation data which can be gathered for each can be visualized as follows:

TYPES OF EVALUATION DATA

	FORMATIVE (Interim)	SUMMATIVE (End of Cycle)
PRODUCT DATA (Learner Changes)		
PROCESS DATA (Supportive Activities)		
CONTEXT DATA (Learning Environment)		

Thus, formative and summative evaluation may include product, process, and context data, all three of which may be collected during a program cycle or at the end of a given program period.

- Product data focus on the outcomes, results, or products of program activity. The purpose of collecting such information is to measure and assess status and accomplishments at the start, during, and at the end of the program. Sometimes postprogram follow-up is also done. Product data should be related to established program goals and objectives. For example, an end-of-the-year summative evaluation of a pilot career education program for grades 11 and 12, with the goal to develop job interview skills for students, might show that 75 percent of the job-interview performance objectives were successfully accomplished by 75 percent of the students.

- Process data focus on the activities and procedures applied to the achievement of the desired outcomes. The purpose of collecting such information is to provide measurements and assessments which will help determine the effectiveness of the various things done in the operation of a program. Process data make it possible to monitor an activity or program to identify and/or predict procedural difficuities before they loom large. For example, early in the program, the decision maker may wish to know whether the teachers and aides are implementing the program's instructional activities as agreed. Process data gathered with a formative purpose can help keep a program on track. Gathered with a summative purpose, they can help in understanding what really happened in the program, after a key benchmark (i.e., end of year) has passed.

- Context data describe the environment in which the program activities are taking place. This might include facilities, equipment, supplies, rules and policies, class organization, teacher skills and behaviors, attitude and support of the principal toward the program, discipline, and scheduling.

  Context data are useful in making judgments on whether program objectives are feasible. They also serve to identify variables that may keep the program from meeting its performance objectives such as a school principal whose attitude will impede a special program imposed on his school. This would be a serious obstacle to the success of the program and would need to be addressed lest the program fail in a starved environment.

An example of each type of information is presented below:

- Product data: The students in the experimental reading program have shown a mean gain of 10 months for every 6 months of instruction.

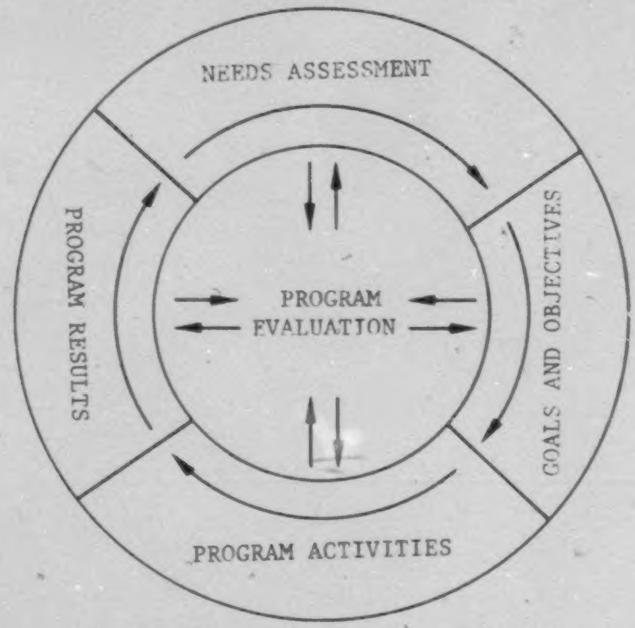
- Process data: The teachers and the aides have carried out all the enrichment program activities as planned.

- Context data: The textbooks arrived two months late resulting in a delay in the implementation of the Career Education Program.

Learning Exercise 1 on page A-25 provides help in understanding types of evaluation data.

Evaluation as an Ongoing Process

Program evaluation is continuing and ongoing. It occurs at the start, during and after a program has been run. One may consider evaluation as the nucleus of the program, for it interacts with the program's needs assessment, its statement of goals and objectives, and program planning and implementation. Evaluation data from the current run of a program becomes needs data for the next run. The entire activity is cyclic, as illustrated below:



The needs assessment gives direction for the development of program goals and objectives; program activities are developed to accomplish the stated program objectives. Carrying them out produces results of one sort or another. The evaluation process interacts with all these stages, and suggests directions for new plans and actions.

The display above suggests that there is constant feedback and revision between program evaluation and each of the four areas. One quickly notes that program evaluation cannot exist as a separate entity; it must be developed as an integral part of the program. Some of the relationships are shown in the following table:

Evaluation Is an Integral Part of a Program		
Why?	Needs Assessment	What needs can you cite that justify the existence of this program?
What?	Program Goals and Objectives	To what needs are the goals of the program related? Of what goals are the objectives of the program a part?
How?	Program Activities	What activities will most likely meet the objectives? How will you plan and carry out activities that will accomplish the objectives?
How Will You Know?	Program Evaluation	What kinds of information should be gathered to determine if the activities are reaching the objectives and consequently meeting the needs?

3. ROLE OF THE PROGRAM EVALUATOR

The role of the program evaluator may be perceived in a number of ways. One view is that of an external person who is called in to assess and verify program success or failure and who will certify to a particular audience that a particular program did or did not attain a specified degree of success. This person may also be seen as an objective and unbiased observer as well as one who may have new insights not readily apparent to those who have been close to the program.

Another view is that of an internal person who is part of the program and whose primary function is to work closely with program staff on evaluation matters. Together they gather information that can be used in improving the day-to-day operation of the program and in learning at the end what happened.

While these two roles are not necessarily mutually exclusive, the emphasis is sufficiently different that the kinds of information they gather and the reports they make are probably significantly different. What one sees as the function of the evaluator is directly related to what one sees as the purposes of the evaluation. Most programs would benefit from an evaluator who is oriented to neither view but is able to incorporate elements of both in his or her evaluation. Some advantages and disadvantages of the internal and external models are listed below.

External Evaluator

Advantages:

1. Probably has more competence in program evaluation techniques.
2. Brings the objectivity of an outside observer.
3. Probably has no vested interest in program outcome.
4. Takes on the major part of the evaluation burden from the existing staff.

Disadvantages:

1. Will take longer to understand a program and the evaluation requirements.
2. Lacks ongoing working relationship with program staff, school and district personnel.
3. Seen as an outsider by program staff.
4. Time schedule of evaluator may not always match local needs.

Internal Evaluator

Advantages:

1. Is apt to be more familiar with the total school setting.
2. Has established working relationship with program staff.
3. Understands channels of communications within the school, the school district and the community.
4. Is familiar with all details of the program.
5. Has a personal interest in the success of the program.

Disadvantages:

1. May have a vested interest in program outcome.
2. May reflect bias of program staff in the design and report.
3. May be overburdened by other duties and unable to devote adequate time to the program evaluation.
4. May not have skills required in evaluation.

4. INITIAL STEPS IN EVALUATION PLANNING

Find Out What the Program Evaluation Is to Accomplish

Program evaluation is frequently thought to be a sequence of activities such as choosing assessment instruments, collecting and analyzing the data obtained by the instruments, and reporting the results. However, good program evaluation consists of much more than that. In developing an evaluation plan, one must ask:

- What are the questions the program planner wishes to have answered?
- What is expected of the program by the different publics, such as students, instructional staff, the principal, the citizens' advisory committee, the superintendent, and the board of education?
- What questions does each of the groups served want answered as a result of the program evaluation?
- What kind of information will each audience accept as evidence for answers to its questions?

Only after addressing such questions can one begin to plan program-evaluation strategies.

Review Needs Assessment, Program Goals and Objectives

One of the first things that an evaluator should do is to become well acquainted with the program to be evaluated. If the evaluation has been part of program planning, the evaluator probably will have been involved in planning the program from its inception. He or she has probably become familiar with learner, educator, and community needs when these were identified. If the evaluator is a more recent arrival, he or she has probably been thoroughly briefed about the program. In situations when evaluation has not been a central part of program plans, an evaluator frequently is not called in until the program has been developed and is about to be implemented. At that time, the evaluator may be provided with a copy of the program document and asked to develop a design and plan for a program evaluation.

If you as an evaluator are faced with this latter situation, be certain that you are provided with the record of previous planning before beginning your task.

As an evaluator, you will want to learn how objectives were set and how they were seen to be related to existing needs. It would be important to document whether all who have an effect on pupil learning were given the opportunity to recommend and to set priorities for objectives to be addressed by the instructional program.

The evaluator's next task would be to review the program goals and objectives to satisfy himself or herself that they relate directly to identified needs and to each other. Part of this task is to see that the objectives are stated in unambiguous terms and that it is uniformly understood what results are expected from the program. This will help the program manager, the program evaluator, and the staff to reach agreement on the direction in which the program should be moving and what it should be accomplishing.

The general responsibility of writing objectives lies with the program director; however, if the objectives are not clear, it is the evaluator's role to seek clarification. At times it may be necessary for the evaluator to assist in the rewriting of the objectives or to suggest alternative statements.

Clearly stated performance objectives are the key to the evaluation process. This is because the objectives set the stage for deciding the nature of the evaluation to be performed, the kind of design to be developed, the types of instruments to be used, the resources required to perform the evaluation, and the style and organization of the evaluation reports. It is imperative that the objectives are carefully formulated, that they relate recognizably to program goals and needs statements, and are framed in ways that make them measurable.

Separate Objectives Statements from Goals Statements

The ability to set meaningful goals and objectives is a very valuable skill which, when applied correctly, will help ensure success in both the program and program evaluation. Goals and objectives can help answer the questions, "What do we want to accomplish?" and "How will we know when we have accomplished it?" A goal is usually a general statement of the long-term results that one might hope to reach or come close to reaching. An objective is developed to reflect a specific outcome of goal-related efforts and is stated in terms of measurable changes or improvements that are expected. The main difference between the two is that an objective by its definition is measurable, whereas a goal is seldom stated in measurable terms.

The following chart summarizes the differences between the two:

GOALS AND OBJECTIVES ARE
NOT THE SAME

Goals	Objectives
• Broad in scope	• Define intent
• General Statements of aspiration	• Define expected outcomes in measurable terms
• Long-term or far-reaching	• Capable of being accomplished within a specified time frame

Determine That Six Components of Performance Objectives Are Present

A well-stated performance objective contains six components that will answer the following questions:

- Who?
- Learns or does what?
- When?
- Under what conditions?
- At what performance level?
- How will it be measured?

The Who relates to the person who performs an activity. The Learns or does what is the activity to be performed. When, Under what conditions, and At what performance level relate to time and performance conditions. How will it be measured relates to assessment techniques.

Summary of Evaluation Planning Stages

The program evaluator begins by reviewing the needs statements and formulations of program goals and objectives. The second step is to review the program activities to determine how they are to meet the stated objectives. If the activities seem not to match the objectives, the evaluator should recommend to the program manager that they be reviewed for possible revision, or that the objectives be changed. The learning exercises at the end of this section are designed to reinforce the information in these pages.

Learning Exercise 2 (A-27) provides help in the identification of measurable objectives. Learning Exercise 3 (A-29) provides help in relating specific objectives to given goals and in rating their relative importance. Learning Exercise 4 (A-39) will be useful in understanding the relationships among needs, objectives, and activities.

5. REQUIREMENTS OF PROGRAM EVALUATION

Key Questions

The ultimate requirement of evaluation is to serve the needs of the audiences to which the program director is accountable. To make certain that both interim or formative and end-of-the-year or summative evaluation are efficient and provide the best possible information for program use, the evaluator should address the following questions early in the evaluation planning stage:

Who requires information? Generally speaking, everyone who has a responsibility for some phase of the program is a decision maker, and he or she will require evaluative information. The evaluator should identify decision makers as early as possible and make personal contact with each to stress the desirability of working closely together throughout the program period and to learn about their information requirements.

What decision-making information is required? During the program period, various staff members with program responsibilities will require information in areas of concern to them. Since responsibilities are not always clear-cut, the decision makers must tell the evaluator which program elements are of interest and what types of information are needed for each. Unless the evaluator has this information, he or she may not be able to provide sufficiently useful data to decision makers. Ideally, to ensure that information collected and analyzed will be as meaningful as possible, the decision maker should formulate questions related to program objectives or concerns that the evaluator should address. Questions submitted by decision makers then can be translated into functional terms for inclusion in the evaluation design, the data-collection instruments, the data-analysis plan, and in the outline for the evaluation report.

When is the information required? Information gathered needs to be both instructive and timely. To assure that program-evaluation reports are submitted when required, a reporting timeline should be developed.

On the following page is a form on which information requirements for program evaluation may be recorded. The evaluator should complete this form as he or she meets with each decision maker to determine his or her information requirements. On page A-19, there is a sample Evaluation Information Requirements Form that has been filled out.

Interim Report ____________
End-of-the-Year Report ________

Program ____________________
Program Director ____________
District ____________________

EVALUATION INFORMATION REQUIREMENTS FORM

1	2	3	4
Who Requires Information?	What Information Is Required?	Date Required?	Use to Be Made of Information

Interim Report ________

End-of-the-Year Report ________

Program ERPHI

Program Director T.H. Collins

District Rosedale

EVALUATION INFORMATION REQUIREMENTS FORM

1 Who Requires Information?	2 What Information Is Required?	3 Date Required?	4 Use to Be Made of Information
Dr. Marie Thomson, Principal	Reports on observations of classroom activities	Nov. 15 Dec. 15 Jan. 15	To determine if the instructional program has been implemented as planned
	Teacher-paraprofessional reactions to inservice training program	Oct. 1	To determine effectiveness of the inservice training program for revising program if required
	Student progress in reading and mathematics achievement	Jan. 15	To determine whether students are achieving at the expected rates

End-of-Year Evaluation

End-of-the-year, or summative, evaluation is critical to decision makers who must decide whether to continue, modify, expand, or discontinue the program. It also serves to identify needs to be addressed in planning the instructional program for the following year.

The end-of-the-year evaluation should answer explicitly those questions that were designed into the evaluation plan at the beginning of the year. It is important, therefore, that the evaluator review the plans for program evaluation prior to program implementation to be certain that all the data which will be required will be available for the end-of-the-year evaluation.

Interim Evaluation

Interim, or formative, evaluation allows decision makers to determine how well program objectives are being met while the program is ongoing and to decide what to do to improve program activities in progress. It is a viable tool for controlling and fine tuning the program.

Care must be taken to evaluate program activities as well as to measure progress towards program objectives. With effective monitoring, deviations from planned activities can be identified immediately and corrected before they adversely affect program outcomes.

Identifying Resources and Constraints

Resources and constraints should be identified during the early phase of evaluation planning. This is important, for it will bring to light the resources needed and those currently available, and will enable decisions to be made on whether to add resources where there are shortages or to get along under constraints. Some resources to consider include:

1. The amount of money budgeted for the evaluation
2. The amount of personnel time available for data collection and record keeping

3. The services available from other agencies (i.e., district, county, and/or state)
4. The instruments currently being administered for other purposes that can provide some of the program-evaluation data

Later in the planning period, when the evaluation plans and procedures have been determined, specific requirements will be identified. A match/mismatch between resources available and those required should be made. As discrepancies are identified, the program director and staff, with the assistance of the evaluator, will be in a position to determine the manner in which the discrepancies can best be handled.

One method of resolving a constraining factor is to change the requirement. Another is to create new ways to meet it. At times, a compromise may be reached with the decision maker as to how much he or she is willing to sacrifice in order to achieve a given feature of a program evaluation. There is also the alternative of eliminating the interim evaluation so that the available resources can be focused upon the longer-range concerns of the end-of-year evaluation. Constraints become evident as planning proceeds. When such problems come to light, the evaluator's advice on alternative solutions is the key to balancing high-level resource requirements against ever-present constraints.

If the program receives categorical funding from state or federal sources, note must be made of the reporting requirements of these agencies. Such external requirements should be combined with local requirements to define the total evaluation requirement. Data for both purposes should be collected and treated as a unit. Duplication throughout the evaluation effort, such as double-data collection, should be avoided at all cost.

SUMMARY

To determine evaluation purposes and requirements, the evaluator:

1. Reviews program records of outcomes of previous program implementation.
2. Determines how the learner, educator, and/or community needs were identified.
3. Determines that program goals match identified needs.
4. Determines that performance objectives are compatible with program goals and needs statements.
5. Determines that performance objectives are written in measurable terms.
6. Reviews program activities to be certain that they relate to performance objectives.
7. Determines end-of-the-year and interim evaluation requirements.
8. Identifies preliminary evaluation resources and constraints.
9. Develops a composite list of resource requirements and submits the list for staff concurrence and board of education approval.

CHECKLIST OF THE STEPS IN DETERMINING EVALUATION PURPOSES AND REQUIREMENTS

	Step Started	Date Completed
DEFINE EVALUATION PURPOSE		
• Determine from decision makers the purpose of the evaluation. The purpose will dictate the types of evaluation that must be conducted.		
REVIEW NEEDS ASSESSMENT, PROGRAM GOALS, AND PROGRAM OBJECTIVES		
• Determine whether a needs assessment was conducted by the decision makers.		
• Review program goals to determine whether they address the needs or problem areas.		
• Review performance objectives to determine that they are compatible with program goals. Are the objectives stated in unambiguous terms?		
REVIEW PROGRAM ACTIVITIES		
• Review program activities to judge whether they can be expected to contribute to achievement of the objectives.		
• If the activities do not match the objectives, recommend that activities or objectives be revised.		
IDENTIFY EVALUATION REQUIREMENTS		
• Request that decision makers identify the information they will require to make end-of-the-year decisions about the program.		
• Determine process, product, and context data that should be collected.		

CHECKLIST OF THE STEPS IN DETERMINING EVALUATION PURPOSES AND REQUIREMENTS

	Step Started	Date Completed
IDENTIFY EVALUATION REQUIREMENTS (cont'd)		
• Determine the information required by decision makers to make interim decisions.		
• Determine when the information is required.		
IDENTIFY EVALUATION RESOURCES AND CONSTRAINTS		
• Determine the resources and constraints which will affect the conduct of the evaluation.		
• Advise decision makers of those resources which are available and those that are required.		
• Submit recommendations to decision makers for reconciling discrepancies between resources available and those required.		

LEARNING EXERCISE 1: TYPES OF EVALUATION DATA

Directions: Classify the following examples by type of evaluation data to be collected.

Types of Evaluation Data

Product -- PT
Process -- PS
Context -- C

Write the abbreviations for these types in the spaces provided.

Example I: The end-of-the-year evaluation indicated that three of the four program performance objectives were met. ☐

Example II: Two of the seven teachers developed their own math materials instead of using those prescribed for the program. ☐

Example III: After the second week of school, all the teachers went on strike. ☐

Example IV: It was determined at the end-of-the-year evaluation that 45 of the 46 instructional activities were implemented as planned. ☐

ANSWERS

Example I:	PRODUCT Explanation: When we speak of performance objectives, we are speaking of learner progress or outcomes.
Example II:	PROCESS Explanation: In this example, the evaluator is looking at the activities implemented to support learner progress or outcome.
Example III:	CONTEXT Explanation: The teachers' strike is a condition which interrupted the instructional program design and which could have kept the program from meeting its performance objectives.
Example IV:	PROCESS Explanation: As in Example II, the instructional activities were designed to support the achievement of the desired pupil outcomes.

LEARNING EXERCISE 2: IDENTIFICATION OF MEASURABLE OBJECTIVES

The following are partial statements of performance objectives. Check whether or not these statements are written in terms that are measurable.

	YES	NO
1. He will be able to understand the principles of citizenship.	_____	_____
2. Each child will be able to name the days of the week in order beginning with Sunday.	_____	_____
3. The students will appreciate the culture of their northern neighbor.	_____	_____
4. The students will construct a log cabin.	_____	_____
5. The teachers will learn the significance of the experience.	_____	_____
6. Children will enjoy going to the library.	_____	_____
7. Parents will become aware of their need to participate in the school program.	_____	_____
8. Students will grasp the concept of good citizenship.	_____	_____
9. Each child will write a composition.	_____	_____
10. The teacher's assistant must know the teaching philosophy of Head Start.	_____	_____

Do the following statements contain the six (6) components found in a performance objective? Be prepared to identify any components that might be missing.

	YES	NO
11. All 2nd grade students receiving remedial math instruction will show a gain of five months in math computations for every five months of instruction. Gain will be measured by the state-approved test.	______	______
12. The students will show a six months' growth in reading comprehension as a result of the remedial reading program.	______	______

LEARNING EXERCISE 3: SELECTING APPROPRIATE OBJECTIVES

This exercise is designed to give you practice in deciding which objectives will best measure progress towards a specific goal. Four different goals are presented. They deal with student achievement; motivation and commitment of students, staff, and community; individualized instruction; and self concept. A list of 11 possible objectives that could be used to assess progress towards the goals is also given. Your task is to take one of the four goal areas and decide which objectives would be most appropriate for evaluating that program goal.

Instructions for Each Panel of Consultants

Suppose that you and other members of your group are consultants selected by the school board to form an advisory evaluation panel. The main task of the panel is to select a set of objectives appropriate to evaluate the goals stated on the "Reading Program Goals Sheet" (A-30). Include only the most important objectives--those you think the district can reasonably afford to pursue and evaluate.

At the end of the panel's meeting, it is expected that the panelists will have:

1. Read the "Brief Description of the Program" and selected one of the four program goals (A-30).
2. Selected from the list of 11 objectives (A-31-33) those most of the members agree are adequate for determining whether the goal selected has been achieved. If the panel members are not satisfied with those listed, they should have developed their own set of objectives (A-34).
3. Individually rated each selected objective using the worksheet (A-36).
4. Tallied the selected objectives according to the ratings assigned by the panel (A-37).

READING PROGRAM GOALS SHEET

Brief Description of the Program

This program is a reading performance contract project funded by the state. The program is located in one of the junior high schools of an urban school district which has shown a great need for special reading instruction. The emphasis of the program is placed on individualized reading instruction. Two main components are: (1) the diagnosis of reading needs of individual students; and (2) the careful planning of reading instruction according to the diagnosis.

Teachers in the program have been given preservice training and will receive inservice training in the use of individualized instruction techniques.

The program is in its first year of operation with contracts between the school district and the teachers working in the program.

Program Goals

Four of the goals of the program stated in the contract are as follows:

I. PARTICIPATING STUDENTS WILL RAISE THEIR READING ACHIEVEMENT PERFORMANCE.

II. ADMINISTRATIVE STAFF, TEACHING STAFF, STUDENTS, AND MEMBERS OF THE COMMUNITY WILL DEMONSTRATE THAT THEY WERE HIGHLY MOTIVATED AND HIGHLY COMMITTED TO A SUCCESSFUL IMPLEMENTATION OF THE READING PROGRAM.

III. INDIVIDUALIZED INSTRUCTION TECHNIQUES WILL BE USED AS THE MAJOR TEACHING STRATEGY IN THE IMPLEMENTATION OF THE INSTRUCTIONAL PART OF THE READING PROGRAM.

IV. PARTICIPATING STUDENTS IN THE READING PROGRAM WILL RAISE THEIR SELF CONCEPTS.

LIST OF READING PROGRAM OBJECTIVES

The planners of the program, together with the designated program evaluator, have developed the following list of program objectives. These objectives are examples which may or may not be appropriate to evaluate one or more of the program goals:

1. All participating students with self concepts below the 30th percentile as measured on a standardized inventory on a pretest will show a gain of at least ten percentile points towards positive self concept as measured by the same instrument at the end of the eighth month in the special reading program.
2. The teaching staff will assess reading skills and design individualized reading activities for each participating student at the beginning and at one-month intervals during the special reading program. The fulfillment of this objective will be measured by the extent of the entries in the locally developed "Student Activities Diary."
3. At the end of the eighth month of the special reading program, at least 95 percent of the teaching staff will have participated in three-fourths or more of the supplementary instructional activities (e.g., inservice sessions, staff meetings) designated in the program. The fulfillment of this objective will be measured by the tallying of attendance in an attendance log.
4. At the end of the eighth month of the special reading program, at least 70 percent of the administrative staff will have participated in one-fourth or more of the supplementary activities (e.g., inservice sessions, staff meetings)

designated in the program plan. The fulfillment of this objective will be measured by the tallying of attendance in an attendance log.

5. Fifty percent of the participating students will show a gain of 15 percentile points or more in reading achievement as measured by a standardized, norm-referenced reading achievement test at the end of the eighth month of the special reading program compared with the results of the same test administered at the beginning of the program.
6. At the end of the eighth month of the special reading program, 70 percent of the participating students will respond correctly to three-fourths or more of the questions on a criterion-referenced test.
7. At the end of the eight-months period, the teaching staff will reflect a measured mean score of 4 or higher on a rating scale with a designated low score of 1 to a designated high score of 5 indicating the extent to which they were personally committed to the successful implementation of the special reading program.
8. When responses are solicited at the end of the eight-months period, participating students will show a measured mean score of 4 or higher on a rating scale with a designated low score of 1 to a designated high score of 5 indicating the extent to which they were personally committed to the special reading program.
9. When responses are solicited at the end of the eight-months period, parents of participating students will reflect a measured mean score of 4 or higher on a rating scale with a designated low score of 1 to a

designated high score of 5 when judging the extent to which they were committed to the successful implementation of the special reading program.

10. At the end of the eighth month of the special reading program, 90 percent of the participating students will have had fewer than three nonjustifiable absences ("cuts"), as measured by attendance records.

11. At the end of the eighth month of the special reading program, 70 percent or more of the participating students will report that they enjoy reading more than they did before entering the program. The fulfillment of this objective will be measured by a locally developed student questionnaire.

SELECTION OF OBJECTIVES
WORKSHEET

Task Number 1: Record the Roman numeral and key words of the goal your panel chose and enter the number of members in your panel at the top of the Individual Worksheet on page A-36. Circle the numbers of the objectives on the Individual Worksheet which your panel feel are appropriate for the evaluation of that goal. If your panel feels that the objectives presented in the list are inappropriate or inadequate, write the objectives your panel agrees are appropriate for your program goal in the spaces labeled PROGRAM OBJECTIVE 12 and PROGRAM OBJECTIVE 13 below.

PROGRAM OBJECTIVE 12

PROGRAM OBJECTIVE 13

Task Number 2: After you have circled the objectives your panel favors or have written in (and circled) new ones in the spaces on page A-34, rate each objective on the following scale: 0 = Not Important; 1 = Important; 2 = Very Important. Put a check in the appropriate boxes on the Individual Worksheet. Next, circle the objective on the Group Report Form, tally the ratings of your panel, and then enter the tallies in the appropriate boxes on the Group Report Form. (On page A-38, there is a filled-out Group Report Form that illustrates how this is done.)

Task Number 3: The leader of each panel reports on the panelists' selection of objectives.

INDIVIDUAL
WORKSHEET

PROGRAM GOAL NUMBER ____________ KEY WORDS ____________

TOTAL NUMBER OF PANEL MEMBERS ____________

Objective Number	Rating of Objectives and Tallies of Ratings		
	Not Important	Important	Very Important
	0	1	2
1	☐	☐	☐
2	☐	☐	☐
3	☐	☐	☐
4	☐	☐	☐
5	☐	☐	☐
6	☐	☐	☐
7	☐	☐	☐
8	☐	☐	☐
9	☐	☐	☐
10	☐	☐	☐
11	☐	☐	☐
*12	☐	☐	☐
*13	☐	☐	☐

---

*These numbers represent objectives which have been written by the panel.

GROUP REPORT FORM

PROGRAM GOAL NUMBER ____________________ KEY WORDS ____________________

TOTAL NUMBER OF PANEL MEMBERS ______________

Objective Number	Rating of Objectives and Tallies of Ratings		
	Not Important	Important	Very Important
	0	1	2
1	☐	☐	☐
2	☐	☐	☐
3	☐	☐	☐
4	☐	☐	☐
5	☐	☐	☐
6	☐	☐	☐
7	☐	☐	☐
8	☐	☐	☐
9	☐	☐	☐
10	☐	☐	☐
11	☐	☐	☐
*12	☐	☐	☐
*13	☐	☐	☐

*These numbers represent objectives which have been written by the panel.

GROUP REPORT FORM

PROGRAM GOAL NUMBER II KEY WORDS HIGHLY COMMITTED

TOTAL NUMBER OF PANEL MEMBERS 7

Objective Number	Rating of Objectives and Tallies of Ratings		
	Not Important	Important	Very Important
	0	1	2
1	☐	☐	☐
2	☐	☐	☐
3	☐	☐	☐
[✓ 4]	☐	2	5
5	☐	☐	☐
6	☐	☐	☐
[✓ 7]	1	6	☐
8	☐	☐	☐
9	☐	☐	☐
10	☐	☐	☐
[✓ 11]	1	2	4
*12	☐	☐	☐
*13	☐	☐	☐

---

*These numbers represent objectives which have been written by the panel.

LEARNING EXERCISE 4: MATCHING NEEDS STATEMENTS TO PROGRAM OBJECTIVES AND PROGRAM ACTIVITIES

In the table on page A-40, you will find three columns. Column 1 contains needs statements; column 2, program objectives; and column 3, program activities. In the spaces provided at the left of columns 2 and 3, write in the numbers of the needs statements that match the objectives and activities.

NEEDS STATEMENTS	PROGRAM OBJECTIVES	PROGRAM ACTIVITIES
1. Parent and community involvement activities were minimal except in the two classrooms where Head Start pupils were enrolled. Since the goals call for extensive community involvement in planning instructional activities, there is a need to increase such participation. 2. Many Head Start-K-3 pupils have limited vocabularies and limited ability to process and verbalize information in standard English. Since the Head Start Program places major emphasis on language development, the staff recognizes a need to upgrade pupils' verbal skills. 3. It has been observed that the individual needs of the target pupils are not being addressed in the classroom. The school philosophy stresses individualizing programs for each pupil. There is a need to insure that the staff implements this philosophy. 4. Almost 80 percent (79.9) of the K-3 pupils tested scored below Q_2 on the total mathematics of the C.T.B.S. given in October 1974. These scores do not appear comparable to pupils' demonstrated ability in other areas. There is a need to identify means of improving pupil performance in math.	--A 5 percent random sample of Head Start and K-3 pupils who have been in the activities of the Oral Language Component will significantly increase (at the .05 level of confidence) scores on the Verbal Expression subtest of the Illinois Test of Psycho-Linguistic Abilities from the October 1974 pretest to the May 1974 posttest. --The median percentile rank for those pupils continually enrolled in K-3 mathematics programs for 125 days or more will be 5 points higher in May 1975 than the median percentile rank previously attained by the same pupils on the pretest in October 1974, as measured by the total mathematics scores from the C.T.B.S. --During the 1974-75 school year, all teachers and aides participating in the special program will provide individualized instruction to the target pupils in their class. Evidence of this will be by periodic classroom observations obtained by the project staff and the evaluator.	--Grouping within each K-3 classroom will be flexible, changing frequently in accordance with the changing needs of the pupils. Groups will be formed by needs identified by diagnosis and recorded on the class profile chart. --All Early Childhood Education Teachers and aides will participate in a four-week summer program during July 1975. The teachers and aides will be involved in problem solving, developing materials, and inservice for improving teaching skills in individualized instruction. --Head Start-K-3 teachers will shape the instructional program in language development by a hierarchy of performance objectives. This activity will be evidenced by a class profile chart covering each of the performance objectives in the classroom. The class profile chart will be monitored in November, February, and May by an evaluation committee made up of two teachers, two parents, the principal, and the Director of Compensatory Education.

PROGRAM EVALUATOR'S GUIDE

Section B

DEVELOP AN EVALUATION PLAN

The Evaluation Improvement Program

PRECIS

The evaluation of a program, like the program itself, operates on a careful plan based on what the program is designed to accomplish and whether it has achieved those results. Thus, having learned the program's objectives and needs (step 1), the evaluator then constructs a series of questions that will be used after the program's completion to determine if it has achieved its objectives. The evaluator then decides upon the instruments that will be needed to gather the necessary information and whether they will be selected from available instruments or developed. A schedule is then drawn up for the administration of the instruments. Decisions must then be made about the kinds of analyses that will be performed on the data and how the final information will be reported and to whom. Finally, to improve the chances that the report will be used productively, the evaluator discusses its uses with all the recipients.

CONTENTS

Page

1. THE PRELIMINARY WORK B-1

    Needs Assessment and Program Goals and Objectives B-1
    Evaluation Design B-1
    Assessment Instruments B-3
    Administration Dates and Personnel B-4
    Data-Analysis Techniques B-5
    Monitoring Program Activities B-5
    Monitoring Dates and Personnel B-5
    Key Reporting Dates B-5
    Who Is to Receive the Report(s) B-6
    Determining How the Data Reports Will Be Used B-6

2. REVIEW OF PRELIMINARY PLANS B-6

3. EVALUATION TIMELINE B-7

4. DETERMINING AND OBTAINING REQUIRED RESOURCES B-7

LEARNING EXERCISE 5: PLANNING FOR ASSESSMENT OF PROGRAM OBJECTIVES B-11

1. THE PRELIMINARY WORK

The preliminary work in developing a general evaluation plan requires consideration of all of the steps involved in the evaluation process. In this section, we shall look briefly at each of the several evaluation steps and identify some of the questions that must be formulated at each. The subsequent sections (C-H) of this Guide will develop in more detail the things that you must know and do as you plan and implement your program evaluation.

To help you visualize all the steps in a total plan, a Program Evaluation Planning Form is shown on page B-2. This form was designed to guide your thinking as you plan for the evaluation of a particular program. The text that follows relates directly to each of the rows on the form.

Needs Assessment and Program Goals and Objectives

The assessment of needs and the setting of program goals and objectives are parts of the program-planning cycle. Since the evaluation of a program is based on what the program is trying to accomplish, the objectives need to be sufficiently explicit so that whatever progress has been made towards reaching those objectives can be assessed. The program evaluator must be involved at this early planning stage, at least to the extent of reviewing plans and making suggestions to ensure that it will be possible to evaluate program objectives.

Evaluation Design

Evaluation design is essentially a systematic approach to the task of gathering information to answer questions or make decisions. The technical part of considering a design cannot begin until some assumptions are made about what the evaluation is to accomplish. Some questions may relate to progress towards program objectives, relative effectiveness of different programs, or relative standing of various groups within a given area. Specific decisions may be made as to keeping, expanding, or dropping a

PROGRAM EVALUATION PLANNING FORM

Program ________________________

Purpose(s) of Evaluation ________________________

Audience(s) for Evaluation ______________________

PROGRAM OBJECTIVES	
EVALUATION DESIGN	
ASSESSMENT INSTRUMENTS	
ADMINISTRATION DATES AND PERSONNEL	
DATA-ANALYSIS TECHNIQUES	
MONITORING PROGRAM ACTIVITIES	
MONITORING DATES AND PERSONNEL	
KEY REPORTING DATES	
WHO IS TO RECEIVE THE REPORT(S)	
DETERMINING HOW THE DATA REPORTS WILL BE USED	

program, adopting or adapting procedures or policies, or where particular programs seem to operate most effectively. The purposes identified in Section A will help determine the important questions to be addressed.

Assessment Instruments

The selection of appropriate assessment instruments depends on the kind of information needed to answer the questions posed in the evaluation. There are a number of different kinds of assessment instruments that are available and that should be considered when a selection is to be made. The following are among the types of instruments to consider:

Norm-referenced tests
Criterion-referenced tests
Questionnaires
Interview guides
Observation record blanks
Rating sheets
Log sheets
Record summary forms
Structured narrative reports

Each of the several kinds of assessment instruments has its own strengths and weaknesses and should be considered in the light of criteria developed for that specific evaluation. Some general criteria might be:

- Does the instrument adequately measure what you want to measure?
- Will the instrument yield consistent results at different times and with different groups?
- Is the instrument appropriate for the particular population in question?
- Is the instrument easy to administer and score?
- Is the cost of the instrument, its administration and its scoring, reasonable and within the budget?

Administration Dates and Personnel

Once dates have been set for administration of the instruments and personnel have been assigned to administer them, there are a number of questions to consider such as:

- Will the assessment dates conflict with other events in a way that might diminish the reliability of the data?
- Will the assessment dates allow for adequate measurement of the program or program elements?
- Will the assessment dates allow the data to be collected, analyzed, and reported to the recipient on time?
- Who can do the assessment with the greatest accuracy and with the least disruption to the regular school schedule?
- Will special inservice training be required to get good results?
- Do individuals involved in the data collection have a vested interest in the outcome?
- Are personnel available on the staff or will outside personnel be required?

Data-Analysis Techniques

Data analysis consists of organizing a quantity of data so that its meaning may be understood. Techniques of analyzing data range from a simple rank ordering of scores to very complex statistical treatment. Data-analysis techniques allow the reader to identify relationships that are not apparent in the initial raw data and make it possible to do such things as compare different groups or the same group at different times.

Listed below are some methods of arranging or displaying data for analysis:

Total raw scores
Mean (average) scores
Median scores
Modal scores
Percentage scores
Rank-order listings
Frequency distribution
Correlations

Monitoring Program Activities

Determining how a program is being conducted is done through a process of program monitoring. Monitoring enables the evaluator to identify unexpected situations or conditions that might impede the implementation of the project and that need the immediate attention of the project director; to collect data for interim reporting, and to observe unanticipated behavior.

Programs frequently include a great many activities. Consequently, it may not be feasible to monitor every one of them, and criteria will have to be established so that the evaluator will monitor only those activities that will yield the most useful information. One such criterion might be the degree to which the program would be affected if a particular activity were or were not continued. Another might be to emphasize activities that appear to be most closely related to the stated program objectives.

Monitoring Dates and Personnel

The same general considerations should be addressed here that were covered under Administration Dates and Personnel.

Key Reporting Dates

In establishing reporting dates, the evaluator must determine when the information in the report is needed by the recipient. Also to be considered is when the information to be reported will be available.

Who Is to Receive the Report(s)

An evaluation report may be designed to answer questions and provide information to a number of audiences. To determine who should receive a report, the evaluator must know the original purposes of the evaluation and what its uses will be, which should have been a part of the initial planning. In any case, a distribution list should be reviewed with the appropriate administrator so that the audiences and intended uses may be verified and the number of copies of the evaluation reports determined.

Determining How the Data Reports Will Be Used

There are several things the program evaluator can do to improve the chances that evaluation reports will be used in a productive manner, and these are discussed in Section G on reporting. For example, the evaluator determines from each recipient of the report the kinds of information needed from the report or the kinds of information the recipient would accept as evidence in regard to a particular question. Joint planning with recipients of the evaluation report is perhaps the most positive action an evaluator can take to insure that the report will serve its purpose.

2. REVIEW OF PRELIMINARY PLANS

Once the preliminary evaluation plan has been completed, it should be reviewed by as many as possible of the people who are involved in the program or who may be affected by the results. In this way, any errors or misconceptions in the plan can be immediately changed or corrected. This review process may bring out honest differences of opinion as to how the evaluation should be designed and implemented. It is not always possible (or even desirable) for a plan to receive universal approval in this kind of review, but a careful reading at this stage may avoid unexpected opposition at the time of reporting.

Review Evaluation Plan with:

- Program staff
- School principal
- District administrator
- Representative from funding agency
- Others?

3. EVALUATION TIMELINE

Following the review of the evaluation plan and the determination that the required resources are available, some sort of implementation schedule should be prepared. One of the more common procedures is to use a timeline on which all actions are listed and the estimated amount of time and actual dates of implementation are recorded. An example of such a timeline, which allows space for an estimate of the amount of staff time required for each of the tasks, is shown on the following page.

4. DETERMINING AND OBTAINING REQUIRED RESOURCES

At this stage of planning, the evaluator must have answers to two questions: (1) What resources are required for carrying out the planned evaluation? (2) Can they be obtained? If the resources do not match the requirements, some adjustments and compromises must be worked out.

The first thing the evaluator needs to identify is the anticipated workload on available personnel. To do this, he must go back to the timeline and look at the estimated total person days for each personnel category. Then he must identify the persons available to fill those needs. For example, if the evaluator has estimated that 300 person days of teacher assistance will be required to carry out the plan, and there are 10 teachers in the same program, that averages out to 30 work days per teacher. If this is a realistic amount of time to expect from classroom teachers and if the teachers themselves agree to the time commitment, there is no problem. If the work load is not acceptable, a more realistic amount of time must be planned.

As staff needs are clarified, there will be several decisions to be made such as: Can the work day requirements be shifted from one personnel category to another? Can additional persons be hired? Can some of the required work days be cut back? Resolution must be reached between work-day requirements and personnel available to fill the requirements.

PROGRAM EVALUATION TIMELINE

														Number of Person Days							
Tasks	Completion Date	1	2	3	4	5	6	7	8	9	10	11	12	Program Evaluator	Program Director	Teaching Staff	Other Professional Staff	Principals	Teacher Aides	Clerical Staff	Other
									Total Person Days											

The costs of required evaluation materials and equipment have to be identified and matched against both available materials and equipment and the budget. Resolution must be reached when there is a discrepancy between required materials and equipment and the available resources.

Costs of such required services as consultants and data processing must be matched against resources. Resolution between required services and resources must be made before the evaluation plan can be put into operation.

In summary, evaluation planning brings about a balance between the resources that are required and those that can be made available. Give and take is involved here: In some cases, resources can be added to match the requirement; in others, the requirement is modified to a less ambitious approach.

CHECKLIST OF THE MAJOR STEPS REQUIRED IN DEVELOPING AN EVALUATION PLAN

	Check		Date Completed
	In Progress	Completed	
• Review needs assessment and goals and objectives to determine their interrelatedness.			
• Identify the purposes for which the evaluation is being conducted and the probable uses of the evaluation.			
• Review objectives to ensure they are written in measurable terms.			
• Identify the questions that must be answered at the end of the year as indicated by the objectives, the purposes, and the probable uses of the evaluation.			
• List appropriate kinds of instruments to gather the information required to answer the questions formulated above.			
• Determine approximate dates when the various kinds of information would most appropriately be gathered.			
• Determine types of data-analysis procedures that would give the most appropriate information to answer the questions formulated earlier.			
• List the activities that need to be monitored together with most appropriate dates to secure the information.			
• List the kinds of reports that will be made, both interim and summative; who will receive these reports; and the dates the reports are due.			
• For each report, list the potential uses to be made of the information and be sure that they match the information to be gathered.			

LEARNING EXERCISE 5: PLANNING FOR ASSESSMENT OF PROGRAM OBJECTIVES

In the Rosedale School District, one elementary school has had particularly rapid growth, resulting in a cultural mix of pupils it has never before experienced. There have been many discipline problems and fights on the school grounds, and the parents have demanded that some action be taken. There have also been complaints regarding the quality of instruction and the achievement levels of the pupils at several grade levels.

Rather than addressing these complaints as isolated problems, a general assessment was made of the educational needs of the total district. As a result of this effort, a number of general goals and specific objectives were developed for four programs, and several changes were planned for implementation during the following year. Because of limited resources, the progress made toward reaching all of the objectives for each of the four programs could not be evaluated during the first year.

Here is some information about the four programs:

Kindergarten Program

At the kindergarten level, there was a high rate of transiency. The staff felt that the test score means of the total kindergarten population were unduly influenced by transients and that this influence distorted the test results downward. The parents wanted to know what a reasonable expectation might be for pupils who attended school on a regular basis, and the staff needed more precise information about all of the pupils' level of achievement.

Of the several objectives for the kindergarten program, the following one was selected for evaluation:

> Kindergarten pupils with an attendance record of 75 percent or better will show an improvement in language skills by achieving a median gain of 30 raw score points on the school-adopted language development test.

Citizenship Program Grades 1, 2, and 3

To deal with the problem of fighting and discipline, a schoolwide program of instruction in multiculture appreciation was developed with a specified curriculum to provide each pupil with planned experiences in a different culture. This was supplemented with a system of counseling in which each pupil participated in a number of different groups.

The objective to be evaluated:

In grades 1, 2, and 3, where all pupils receive group counseling and instruction in the appreciation of multicultural differences, pupils will demonstrate an improved knowledge of the cultural differences emphasized in the curriculum as evidenced by the district-made test covering the subject matter taught.

The incidence of fighting on the playgrounds will show a 20 percent reduction as compared to the records of the previous year.

Mathematics - Grade 10

As one result of a community survey, a minimum proficiency level in math was established for all grade 10 pupils, and those not meeting this level were enrolled in a remedial math class where individualized instruction was to be emphasized and individual diagnostic records were to be maintained.

The objective to be measured:

Tenth grade pupils receiving remedial math instruction will show a five-month mean gain in math computation for every five months of instruction as measured on the school-adopted standardized math test.

English - Grade 12

All high school seniors were required to take at least one semester of English. Many of the graduates who went on to colleges or universities, however, were not passing the test for written expression, even though they had taken the college preparatory course. The school board directed that standards be developed for the class and that an evaluation of the results be made.

The objective to be measured:

All high school seniors receiving a grade of C or higher in senior English writing class and making application to a college or university will earn a passing score on the writing section of that institution's entrance examination.

Using the Program Evaluation Planning Form on the next page, you may assume that an adequate needs assessment has been done, the goals have been adopted, and that the programs are properly designed to meet the identified needs.

Your group is to select one of the four programs and fill out the questions in each column on the planning form. The questions on page B-15 may be helpful in filling out the form.

PROGRAM EVALUATION PLANNING FORM

Program ________________________

Purpose(s) of Evaluation ________________________

Audience(s) for Evaluation ______________________

PROGRAM OBJECTIVES	
EVALUATION DESIGN	
ASSESSMENT INSTRUMENTS	
ADMINISTRATION DATES AND PERSONNEL	
DATA - ANALYSIS TECHNIQUES	
MONITORING PROGRAM ACTIVITIES	
MONITORING DATES AND PERSONNEL	
KEY REPORTING DATES	
WHO IS TO RECEIVE THE REPORT(S)	
DETERMINING HOW THE DATA REPORTS WILL BE USED	

PROGRAM EVALUATION PLANNING FORM

Program ______________________
Purpose(s) of Evaluation ____________________
Audience(s) for Evaluation ____________________

PROGRAM OBJECTIVES	What objective is being evaluated? What is the goal or need statements to which this objective relates? Is this objective written in such a form that it can be measured? Is the implied measure appropriate for the objective?
EVALUATION DESIGN	What questions must this design address? What information must this design be able to produce in order to answer these questions? To what purposes of evaluation do these questions relate? What information will the audience accept as evidence related to the purpose of the evaluation?
ASSESSMENT INSTRUMENTS	What kinds of assessment instruments will be most appropriate to secure the information required in the design? (Norm or criterion referenced tests - questionnaires - interviews - observations - rating scales - log sheets - narrative reports)
ADMINISTRATION DATES AND PERSONNEL	During what month or months should assessment take place? Who would be the most appropriate person to collect the data? Who is responsible for assigning personnel and dates?
DATA-ANALYSIS TECHNIQUES	What kinds of scores will be most useful in providing the information needed, as identified in the purpose and in the design? What kinds of data analysis will be most appropriate? Will outside help be required to do the required analysis?
MONITORING PROGRAM ACTIVITIES	What activities are central to the accomplishing of the objectives? What information must be collected to accomplish the purposes of the evaluation?
MONITORING DATES AND PERSONNEL	Who will perform the monitoring function? How frequently must the activities for this objective be monitored? To whom should the monitoring be reported?
KEY REPORTING DATES	Who will be interviewed to ensure that reporting dates meet decision or user requirements? Who will establish reporting deadlines?
WHO IS TO RECEIVE THE REPORT(S)	What different audiences will receive evaluation reports on this objective? Have the questions identified by the audiences during the initial design step been addressed in the evaluation report? Have the purposes of the evaluation been accomplished?
DETERMINING HOW THE DATA REPORTS WILL BE USED	What activities have been planned to ensure the most effective use of the evaluation reports?

PROGRAM EVALUATOR'S GUIDE

Section C

DETERMINE THE EVALUATION DESIGN AND DO THE SAMPLING

NOTE TO USERS OF THIS *GUIDE*

In several sections of this *Guide*, as in this one, each Learning Exercise is placed immediately after the discussion of the topic(s) the exercise is based on. In the other sections, the Learning Exercises are grouped at the end. There are advantages in both ways of presenting these materials. Which way do you prefer?

The Evaluation Improvement Program

PRECIS

Evaluation design is the key to obtaining valid and reliable information for decision making, which is, of course, the most important purpose of program evaluation. Applying the principles of design helps assure a high level of objectivity by eliminating personal opinion as a major factor in program evaluation. Good design enables one to compare with confidence the performance of students on such dimensions as past and future, program groups and nonprogram groups, Treatment A group and Treatment B group. It lets us know what change has occurred, what the gains or losses have been, which of several procedures is preferable, and whether sought-after objectives have been reached.

If knowing about gains or growth is critical, a pretest-posttest design should be selected. If assurance is needed that program treatments have had impact, nonprogram groups should be measured along with program groups. If it is necessary to look at subparts of the program for cause/effect relationships, or at various combinations of participants, an expanded factorial design should be chosen. If homogeneity across groups is an important factor, individuals should be assigned to groups on a random basis.

To be effective, the design of an evaluation plan should be selected well in advance of program activities. Advance planning of this sort is not only more effective but also more economical, for it allows the use of sampling procedures the evaluator can use to apply the design to representative segments of students, of other populations, or of instruments.

In summary, careful attention to design will help increase the confidence you have in the results of a program evaluation. Careful attention to, and use of, sampling will produce comparable results using considerably fewer subjects in the program-evaluation activities.

CONTENTS

Page

1. INTRODUCTION TO EVALUATION DESIGN C-1

2. HOW TO SELECT A DESIGN C-1
   At What Level Are the Students Functioning? C-2
   How Much Growth Occurred During the Program? C-3
   How Does This Growth Compare with Expectations? C-5
   What Elements of the Program Contributed to the Gain? C-8

3. INTERPRETABILITY OF RESULTS C-12

4. WHAT TO DO TO AVOID PITFALLS C-13

5. OTHER CONSIDERATIONS C-14

6. MONITORING ACTIVITIES C-15

LEARNING EXERCISE 6: EVALUATION DESIGN C-16

7. INTRODUCTION TO SAMPLING C-20

LEARNING EXERCISE 7: SAMPLING CONSIDERATIONS C-24

8. HOW TO SELECT A RANDOM SAMPLE C-26
   Check Your Random Sample before Collecting Data C-27
LEARNING EXERCISE 8: RANDOM SAMPLING C-28
   Stratified Random Sampling C-32
   Matrix and Multistage Sampling C-34

9. HOW LARGE SHOULD A SAMPLE BE? C-35

10. TWO CASE STUDIES ON SAMPLING C-37
   Number 1 . . . . C-37
   Number 2 . . . . C-39

11. A FINAL WORD ON DESIGN AND SAMPLING C-40

1. INTRODUCTION TO EVALUATION DESIGN

A question of primary importance in program evaluation is: "How much more did pupils learn by participating in the program than they would have learned without it?" The answer involves two bits of information:

- How much students improved between the time the program began and ended
- An estimate of how they would have done without the program

The first is relatively easy to answer if proper instruments are used. The second is more difficult.

The adequacy of the design the program evaluator selects can be judged by the extent to which the results can be interpreted and generalized to other similar kinds of groups and programs. An adequate design helps raise the confidence the evaluator and program director can place in the results.

2. HOW TO SELECT A DESIGN

The particular design you decide upon will depend upon the types of questions you want to answer. Most program evaluation has, in past years, been more subjective than objective. Have you ever heard someone say, "Of course it's a good program. You can just tell by observing the students and the teacher in action. Anyone can tell it's good."?

> I know it's good because I feel it here.
> program ⟶ warm feeling.

This is not what program evaluation is all about. What is needed is the kind of design that will serve to provide valid data for decision making and sound justification for continued funding.

There are many designs that are usable in a school setting that do that, but only a few will be presented here. They deal with these questions:

1. At what level are the students functioning at the beginning and at the end of the program?
2. How much growth occurred during the program?
3. How does this growth compare with our expectations?
4. What elements of the program contributed to the gain or loss?

At What Level Are the Students Functioning?

Knowing where students are functioning at the outset is necessary to successful implementation of the program. A program may be designed for certain types of students with skills at an assumed level. Unless data are available that confirm the fact that pupils in that program really are at that skill level, you may miss the target. A test given early in the program can give you this information.

Are the students entering this program really functioning at the level expected?

premeasurement ⟶ program

This is a legitimate use of a single testing session early in the program. It provides a benchmark that indicates where things stood at the beginning.

Sometimes the program is well under way before an evaluation plan is developed, which is a practice to be discouraged. What does the evaluator do in such a case? One option is to plan a design that uses a test at the end of the program.

At what level are students functioning at the end of the program?

program ⟶ postmeasurement

If the design provides only for this, it really tells us nothing about the effectiveness of the program. It tells us something about what students know after the treatment of the program, which may be used in exploring and developing ideas for further program planning. But it is not a design that will lead to any useful conclusions about the effectiveness of the program at hand.

Fortunately, there are some "retrospective" measures that can be attempted. Unless the year is about over, a test given even as late as midway in the program cycle will give some information against which final results can be compared. Another useful option may be to go to whatever student records there are that indicate general levels of past performance. The conversion of historical data to baseline data is, at best, "messy," but it does offer an expedient measure that can save a late-starting program evaluation.

How Much Growth Occurred During the Program?

Only when this question is asked do we begin to be concerned about the effect that the program has had on the students.

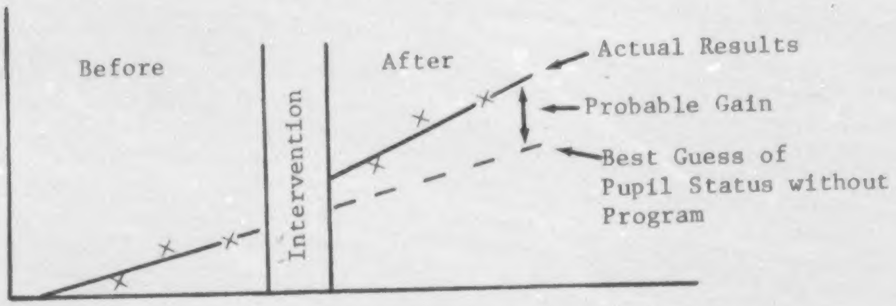
How much did students improve
during the program?

premeasurement → program → postmeasurement

While this design will tell us how much change has occurred, it, too, really does not address the basic question, "How much more did pupils learn by participating in the program than they would have learned without it?"

A single-group time-series design uses students in the program as their own control group. The same measurement is made on the same students at regular intervals several times before and after the program. If the program appears to disturb the trend of measurement results in a positive way, this may be evidence that the program has been effective. This design might be used within a program if a teacher wished to increase the number of new words learned each week beyond the current rate and introduced a special reward system for the child learning the greatest number of new words each week for the next month. Weekly records kept before and after introduction of the system might look like this:

Example:



This design addresses the same question, but gives better evidence as to whether a real change has occurred.

How much did students improve during the program?

premeasurement 1 → premeasurement 2 → premeasurement 3 —

→ program —

→ postmeasurement 1 → postmeasurement 2 → postmeasurement 3

How Does This Growth Compare with Expectations?

Usually the reason a new program is instituted is that it is thought to be better than what currently exists. To find out if the program is better, the evaluator needs to consider what happened as a result of the program and what might have happened if the program had not been introduced.

Whenever possible, some kind of reference group should be used to compare the status of program participants to that of similar persons not participating in the program. There are three kinds of reference groups: (1) control groups, (2) comparison groups, and (3) norm groups. Care must be taken that persons in the reference group are as much like those in the program as possible with respect to age, ability, reading level, ratio of males to females, number of minorities, etc. Persons need not match on a one-to-one basis (in fact, it is better they do not), but the overall group profiles should be similar. If it is not possible to find a reference group in the same school, try to find one in another school which is similar. In determining whether another is similar, consider the following:

- prior achievement of students
- population density of community
- size of school
- school organization (e.g., K-3 vs. K-6)
- teacher training and experience
- median family income
- expenditures per student
- eligibility for state and federal programs
- racial composition
- administration/teaching philosophies

Program and control groups are formed when pupils (or other participants) are randomly assigned to program activities or nonprogram activities within a program-evaluation model. If it is possible to make random assignment to one group or the other, the concerns that program versus nonprogram groups be alike do not apply. Even so, it would be a good idea to check with a random assignment to see whether there may be drastic differences between groups. Randomization is an acceptable way of assuring "likeness." If it is administratively feasible, it is the best possible way to control grouping.

Random
Assignment

How do nonprogram students identical with
program students compare with regard to growth?

program students: premeas. → program → postmeas.
nonprogram students: premeas. → postmeas.

Sometimes in such a design, it may be advantageous to plan for no pretest. This may be because new and unfamiliar concepts are to be taught and it seems unlikely that information will be gained by a pretest. Or, in the case of very young children, the amount of testing that can be done is very limited. Or sometimes the use of a pretest itself may influence or contaminate the behavior of the students. With random assignment of pupils to program and nonprogram groups, then a posttest-only design will yield the needed information on growth.

Random
Assignment

How do nonprogram students identical with
program students compare with regard to growth?

program students: program → postmeasurement
nonprogram students: postmeasurement

A comparison group is one in which existing classes of comparable pupils have been identified as nonprogram participants. Comparison groups are not specially organized, as through random assignments. Both control groups and comparison groups must be given the same instruments on the same schedule as the program group. If these groups are in the same school with the program group, care must be taken that "contamination" does not occur. All too frequently if a new method in the new program is getting good results, the word will filter through to the control or comparison-group teacher, and the new method will find its way into that group's classroom. If this happens, the entire evaluation may be invalid.

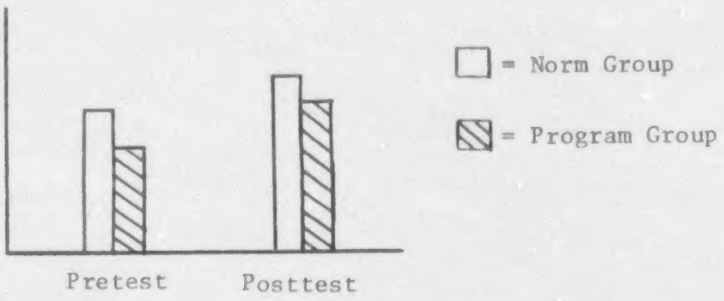
Nonrandom Assignment

How do comparable nonprogram students compare with program students?

program students: premeas. → program → postmeas.
nonprogram students: premeas. → postmeas.

When neither a control nor a comparison group is feasible, norm groups may be used. Norms on standardized tests are the most common example of this type. Since norm groups are generally representative of a broader population, they may not be comparable to the particular program group, and care must be taken when interpreting the findings.

When norm groups are used as a basis for comparison, the expectation is that pupils will be relatively higher with respect to the norm at the end of the program treatment than they were at pretest.



How do program students compare with some norm group?

program students: premeasurement → program → postmeas.
norm groups: norm_1 → norm_2

Another type of comparison that can be made is to compare the progress of a special program group with its own past performance. If students who have typically been growing at a rate of one-half a year for each year in school achieve a year's growth in the first year of a new program, this may be evidence that the program is effective.

How does students' achievement in the new program compare with their past performance?

current program: premeas. ⟶ program ⟶ postmeas.
past performance: average growth per year

What Elements of the Program Contributed to the Gain?

Why is this question important?

- If we knew what elements of the program made more difference and which made less, we would be in a position to make improvements in the program.
- Information about effective and less effective elements could be used to reallocate resources of time and money.
- If we knew what elements were essential, we could predict the chances of the program being successful elsewhere.

Unfortunately, finding out what parts of a program contributed most to the gains achieved is a very difficult task. Effective programs may contain many factors related to success, and the task of separating them makes the evaluation design more complex. Thus, with this kind of design there is an even greater need for planning carefully in advance. One certainly would not construct this type of evaluation design after a program has been started.

Ideally, the evaluator would like to take into account as many factors as possible at one time in order to isolate just what it is that is causing the changes which do occur. Unfortunately, the more factors that are put into a design, the more complex the design becomes. Suppose we want to investigate which of two types of instructional materials can be used as effectively without aides as with aides. This would require at least four classrooms. That is, there would need to be at least one teacher for each combination of aides and type of material -- aides present, type 1; aides absent, type 1; aides present, type 2; and aides absent, type 2.

		Instructional Material	
		Type 1	Type 2
Aides	Present	Teacher 1	Teacher 2
	Absent	Teacher 3	Teacher 4

If the teachers' skills and experience are very similar, this simple design could be used. (In the real world, however, it is very unlikely that your teachers will be so evenly matched in training, experience, and talent.) In order to allow for differences in teachers, at least two should be assigned to each combination of aides and type of material, making an eight-classroom design.

		Instructional Material	
		Type 1	Type 2
Aides	Present	Teacher 1 Teacher 2	Teacher 3 Teacher 4
	Absent	Teacher 5 Teacher 6	Teacher 7 Teacher 8

This design is based on the assumption that other relevant factors are held constant (i.e., amount of instructional time, type of instructional method, ability level of pupils, etc.) across all classes. If pupils can be randomly assigned to classes, factors associated with pupil characteristics should be about equally distributed and should not unduly influence one group more than another.

This type of design is very flexible and can be expanded to include more factors and more than two categories or levels per factor. Obviously, the number of classrooms increases rapidly, and only the larger districts will have enough classrooms to use more complicated designs. An expanded design, however, makes it possible to seek answers to many questions in the course of the program evaluation. This is fortunate in those cases in which it is believed that two or more factors interact with each other: the use of aides, for example, the availability of a reading laboratory, or individualized curriculum materials. In an expanded design, each of the factors may be considered individually or in terms of the effects they have on one another.

Notice that to use an expanded design you must define the factors very precisely. For example, individualized instruction may mean a student moves at his own rate or that he receives individual help or that he has his own objectives. Moreover, you might want to look at the degree, extent, or intensity of the factor. You might compare the effects of individual help every day to the effects of individual help every other day.

There are some complications which should be mentioned. Is the teaching comparable in the several classrooms? Is the design for each classroom followed with rigor? Are staff members committed to their approach? A negative answer to such questions could bias the results.

The questions you must consider in planning this or any other kind of program evaluation design are:

1. What are the most important questions that a program evaluation can help answer and how much information do you need to answer these questions?
2. What other relevant variables are there that might have an effect on the outcome and how can they be held constant or randomly distributed among classes or subgroups?

The second question involves a factorial design. Only larger schools and districts have enough classrooms to use such a design. Under certain conditions, smaller schools may undertake a modification of it.

A factorial design may be used if individual pupils can be randomly assigned to each different combination of factors (each cell in the table on page C-11). For example, to study the effects of differing amounts of time spent in a language laboratory by students at various levels of ability, one might randomly assign pupils to make up a group for each condition. Such a design might look like this:

			Amount of Time Spent in Laboratory			
			10 Min.	20 Min.	30 Min.	40 Min.
Random Assignment	Ability Level	High	8 Students	8 Students	8 Students	8 Students
		Medium	8 Students	8 Students	8 Students	8 Students
		Low	8 Students	8 Students	8 Students	8 Students

In this example, students can be sorted on some dimension of ability, and the amount of exposure given each student can (and would need to be) carefully controlled. Only factors that can be managed in this manner can be studied using this kind of design.

While designs as extended as this have not been commonly used in program evaluation, they are very powerful and have considerable potential for situations in which the use of a reasonably sophisticated program evaluation is critically important.

3. INTERPRETABILITY OF RESULTS

Whether results of measurement within a given design have significance depends, in part, on the extent to which the outcomes are the result of treatments and not of some other cause or combination of causes. Some nontreatment causes might be the following:

1. Differential drop-out or attrition rates in groups being compared. Even though you may exclude from your analysis those students who have not been in attendance during the entire program, the transfer of three of your "best" students can have a marked impact on your results.
2. Failure to account for some related condition that directly affected the results although it may not have been expected to. For example, if the program requires students to spend an extra half hour at school each day, and one of your objectives is to improve attitudes toward school, you may find that nonprogram students who get to leave earlier will have more positive attitudes.
3. Contamination between program and nonprogram students. This occurs, for example, when your special program teacher becomes enthusiastic about a filmstrip used in the program and lends it to a nonprogram teacher.
4. The Hawthorne Effect. If much to-do is made about a new program, improvements that were not caused by the program may occur because of 1) the novelty, 2) an awareness that one is a participant in a special group, or 3) a new environment which includes observers, special procedures, equipment, and so on. Improvements caused by these influences are usually short-termed and will probably disappear over time.
5. Evaluator (or teacher) bias, the "Self-Fulfilling Prophecy." This is a well-documented human hazard in evaluation design. The evaluator (teacher or other observer) has preconceived ideas of what is going to happen and sees what he or she expects to see, ignoring indications contrary to expectations.

6. Change in school programs, personnel, facilities, class size, community factors and other such conditions. These factors can affect student performance though none may be related to a program, its treatments, activities, and its evaluation.

4. WHAT TO DO TO AVOID PITFALLS

1. Be sure to take into account time spent on a given subject area. Pupils with high absentee rates will affect results. If your comparison group is spending twice as much time as your program group, time alone may prevent your program group from comparing favorably.
2. If you use a norm-referenced test, try to select one whose normative data were collected at about the same time of year (fall or spring) you plan to use the test.
3. Be sure you use the appropriate test level. If most students answer nearly all or hardly any of the items correctly, measurement will be both invalid and unreliable.
4. Be sure pretest and posttest are comparable; it is preferable to use different forms of the same level of the same test. This is not as critical if you have a good control or comparison group. If one of the other designs is used, however, there is no way to compare the results of two tests normed on different groups of students unless those tests have been statistically compared. In some cases, conversion tables may be available, as in the case of eight of the most commonly used reading tests for grades 4, 5, & 6.*

---

*Loret, P.G. et al. Anchor test study: Equivalence and norms tables for selected reading achievement tests (grades 4, 5, & 6). Office of Education Report 74-305. Washington, D.C.: U.S. Government Printing Office, 1974.

5. If students have been selected for a program on the basis of extreme scores (disadvantaged or gifted), do not use the test that produced these scores in your regular program-evaluation testing. You will need tests of less or greater difficulty.

6. Do not lean heavily on grade-equivalent scores for measuring results. Design your study in such a way that raw scores can be converted to standard scores at the data-analysis stage. Grade-equivalent scores are suitable for descriptive purposes (see section on data analysis).

7. Be sure the comparison you make is between program group and comparison group at the start and at the finish. If you have selected a good comparison group and have an effective program, the initial differences between the two groups will not be significant, but the posttest differences will be.

5. OTHER CONSIDERATIONS

Part of sound program-evaluation design is to plan how to determine whether or not a given program would be effective in another setting. This may seem simple until you consider the multitude of factors involved. Think, for example, of the variations you can expect among students, teachers, schools, and communities:

- STUDENT FACTORS - sex, age, attitude, courses taken, etc.
- TEACHER FACTORS - experience, enthusiasm, skills, intelligence, etc.
- SCHOOL FACTORS - size, budget, attitude toward innovation, etc.
- COMMUNITY FACTORS - income levels, parents' occupations, perceived value of schooling, etc.

However, the situation seems simpler again when we realize that not all of these factors are likely to be important to a particular type of program. The problem, then, is to determine which factors are relevant in testing to see whether a program can be successfully exported from one place to another.

The first step is to review a very complete description of the program itself. What is it about the program that is really essential? What are the components that must be transferred to the new location? What characterized the classes with which the program worked well?

Subsequent steps in planning for a possible exportation will carry the program planner through all the stages in the new site that had been pursued in the old one, but under an entirely different set of conditions.

6. MONITORING ACTIVITIES

Whatever design is used, some procedure should be established to monitor the activities to assure that the design maintains its integrity and that the program is being implemented as intended. Are the new materials being used with the appropriate groups? Do not underestimate the cross-fertilization that may take place between teachers using different methods or materials. A close check must be made periodically to determine whether the program the evaluator thinks he is assessing is the one being carried out.

LEARNING EXERCISE 6: EVALUATION DESIGN

A 7th grade special reading program enrolls 100 students in four different classes. Enough new reading materials for two classes have been purchased. The teachers are prepared to implement an individualized approach with a diagnostic/prescriptive series of tests and activities to accommodate varied reading skills. The approach used until now has not provided for individualization and has not used the new materials, but has been fairly successful. The faculty and administration would like to know if the new approach is really any better and which of the new ideas is more beneficial. Your task is to develop an evaluation design to provide them with as much information as possible. There are six different teachers whom you could assign to the four different classes.

1. Plan a design to compare the old program with the new program using program ⟶ measurement notation or a matrix.

2. How would you assure that the groups of students were comparable?

3. How would you assign the teachers in this situation?

4. What would you do to guard against contamination?

5. What would you do to guard against any Hawthorne Effect?

6. What would you do about the amount of instructional time devoted to reading in each of the four groups?

ANSWERS

1. Plan a design to compare the old program with the new program.

Random Assignment

Individual Instruction		New Materials	
		Yes	No
	No	25 Students	25 Students
	Yes	25 Students	25 Students

or

Group 1: premeasurement → new materials and old instruction → postmeasurement

Group 2: premeasurement → individual instruction and old materials → postmeasurement

Group 3: premeasurement → new materials and individual instruction → postmeasurement

Group 4: premeasurement → old program → postmeasurement

2. How would you assure that the groups of students were comparable?

   Randomly assign students to the four different groups. Compare results of pretest. Compare ethnic composition of group, occupational level of parents across groups, and number of boys and girls in each group. If real differences exist on any of these comparisons, interchange students to better balance the groups.

3. How would you assign the teachers in this situation?

   Compare training and experience of the six teachers. Use "soft" data if available (What reputation does each teacher have with his or her peers and students?). Select the four teachers most alike on these variables and assign them randomly to each of the four groups.

4. What would you do to guard against contamination?

   Meet with the teachers to explain how you are planning to study the effects of the new materials and individualized instruction. Seek their cooperation and explain how sharing of either information or materials can destroy the evidence needed to make good decisions.

5. What would you do to guard against any Hawthorne Effect?

   Refrain from making any predictions about the relative merits of the old program, new material, or individualized instruction. Be frank with the teachers; do not give the impression this is some kind of contest. Advise teachers not to confide in pupils that some kind of experiment is going on.

6. What would you do about the amount of instructional time devoted to reading in each of the four groups?

   Presumably, in the 7th grade, classes are of the same length; thus, available instructional time is the same for all groups. If this is not so, arrange schedules so that each group does have the same amount of exposure.

   Absenteeism may, however, occur at different rates in the four groups. Therefore, teachers should be asked to keep attendance records. At the end of the year, all pupils on whom there are both pre- and posttest scores should also have complete records on attendance. Before analyzing tests, attendance rates for the four groups should be compared.

7. INTRODUCTION TO SAMPLING

One of the early decisions in planning is to determine whether data should be collected from the entire population involved in a program or from only a representative part of that population. If only a portion of the population is used, that portion is called a SAMPLE and the process used to select it is called SAMPLING.

Sampling procedures are important because they allow the evaluator to collect information more economically. A sample that is representative and carefully selected permits the evaluator to make inferences, generalizations, and to draw conclusions about an entire population by applying the evaluation only to the sample.

People use sampling in everyday life, often unconsciously. A consumer who samples a quart of milk would not need to drink the whole quart to determine whether or not it was sour. However, sampling people is not as simple as sampling milk. It is, therefore, important to specify the sampling criteria, as these will define the population to which the findings are expected to be generalized.

Sampling is especially appropriate for program evaluation. Most evaluation activity in the classroom is for the purpose of grading individual students. A score in these cases is needed for each individual in the population. Information at the individual level is useful to the teacher and necessary for student assessment. Many school personnel, as a consequence, have become accustomed to thinking only in terms of whole populations. However, program evaluation requires only information about the effects of the program on students as a group, not as individuals. Furthermore, contrary to popular belief, sample size can be comparatively small and still provide reliable information, provided human characteristics known to contribute to variability in responses are used as the basis for sampling.

It is almost certain to be more economical and more effective to select a sample of students in the program and administer data-collection instruments to them.

Most of the advantages of sampling are related in some way to lower costs. It is less expensive to gather and analyze a hundred scores than a thousand scores. Expense is an especially important factor with some types of instruments. One teacher can administer an hour test to 30-50 students at one time within one hour. In the same time he could interview only a few. Time and money may also be saved in scoring, especially with some types of instruments.

As a result of lower costs, sampling may make it possible to use some types of instruments which would not otherwise be feasible. For example, suppose that the ideal method of data collection was by an interview. Gathering data by this means requires not only large amounts of time but also considerable training of those who will do the interviewing. For this reason, an interview instrument might be rejected if data were needed from the entire population. But if a sample is used, interviewing may be feasible.

The type of instrument used is related to the type and number of objectives that must be assessed. Often, the most important objective of a program cannot be measured by a paper-and-pencil test. It may involve an attitude or behavior outside the school setting. Anticipating the time and trouble involved in observations or follow-up studies, the school may decide that an evaluation to assess the success of such a hard-to-reach objective is not realistic. But sampling might enable the school to gather data on a limited number of cases, making it possible to carry out an evaluation of that objective. Likewise, a school may want data on several hundred objectives. A sampling process could be designed whereby different groups of students are assessed on different groups of objectives. The school thus obtains the needed data, but no one student is subjected to exhaustive testing.

The economies resulting from sampling may enable the evaluation program to use more than one measuring instrument for a given objective. For example, a questionnaire may be developed to assess an attitude. But do students' written responses reflect their actual feelings? Supplementary use of interviews, observations, or open-ended questions with a sample of the students might provide a way to validate the questionnaire results. Computation skills can be assessed by a paper-and-pencil test in the math classroom. But the program evaluator might also be interested in whether these skills manifest themselves in a social studies class. Comments from other teachers on a sample of students might help to assess the transfer of those skills to other situations.

Sampling may be used in a variety of ways to provide new data or to improve on data already collected. For example, teacher and course evaluations by students, peers, and administrators are frequently administered at the end of the school year. If such data were collected using samples in the fall, winter, and spring, guides for instructional improvement could be provided during the year.

In considering a sampling procedure, there are a number of preliminary questions which need to be raised:

- Will a sample provide the representativeness which is necessary?
- Will sampling be more efficient than using the total population?

There are other questions. Use of a sample may result in some loss of accuracy in the information obtained for program evaluation because a score from every student in the population may not be available. This may raise the question of what degree of accuracy loss is acceptable in return for the saving of time and trouble. On the other hand, gathering data from a sample rather than from the whole may yield more accurate results. A large amount of data which is carelessly collected is useless. A smaller amount collected under carefully controlled conditions is very useful indeed.

If a sampling procedure is to be used for all or part of the evaluation, the questions of the size of the sample and the ways to obtain it must be handled next. Relatively smaller samples can be used when the population tends to be homogeneous, when larger differences are expected on the factors measured, or when many of the factors contributing to variability are controlled.

Relatively larger samples are appropriate when the population is heterogeneous, if there are many uncontrolled factors, or if the differences in the factors are expected to be slight.

The adequacy of the findings is more likely to be influenced by sample design than by sample size. All of the advantages of sampling are based on the assumption that the sample is representative or typical of the total population. If the sample is not representative of the population, the data obtained will be misleading. If we assess the effects of a reading program from one teacher's class--a sample, but not a representative one--we may have information about the effectiveness of that particular teacher but not about the effectiveness of the program. A sample is representative to the extent that it reflects the characteristics of the overall population in that setting. The technique used to obtain representativeness is random sampling, which is discussed in part 8.

LEARNING EXERCISE 7: SAMPLING CONSIDERATIONS

Read the situation given below and attempt to identify some of the difficulties and possible resolutions.

> A small school has used teacher-parent conferences as a substitute for report cards. The principal was responsible for the innovation and believes it to be successful, but he wishes to have the views of others who are involved in the process. He designs a questionnaire which is placed in teachers' boxes and sent home with students. The returned questionnaires are to be tallied for use in determining whether the conferences should be continued.

Difficulties	Possible Solutions
1.	
2.	
3.	
4.	
5.	

ANSWERS

1. Teachers and parents are only part of the population in question. What about students? The population has not been completely defined.

2. What proportion of teachers and parents will return the questionnaire? Will those who do be representative?

3. The principal is known to favor the use of conferences. Will this influence the number or nature of the responses or their interpretation?

4. It may be advisable to allow all parents and students to express their views. The questionnaire could be made available to everyone with the returns kept separate from those in the sample. Asking everyone to respond will reduce the chance that some people will wonder why they were excluded. Cost factors and local conditions will dictate whether or not this is advisable.

8. HOW TO SELECT A RANDOM SAMPLE

The method of selecting a random sample is not complicated, but there are some common misconceptions as to how it is done. The selection must be done in such a way that each person in the population has an equal chance to be drawn. A commonly used procedure is to select every 10th or 20th name on a list. If the first name is chosen randomly from among the first 10 or 20, every name has the same chance of being included in the sample. Samples differ, however. The list may be alphabetical, it may be organized by grade or age, or it may be totally unorganized, or random.

Using an alphabetical list is easy and usually free of bias unless there are periodic features in the list which coincide with the sampling interval. For example, if some ethnic names tend to group themselves at specific points in an alphabetical list, you could run the danger of undersampling those groups.

Although this method is popular and used widely, using a table of random numbers is a better procedure.*

Excerpt from
A Table of Random Numbers

Row Number				
1	50691	91653	88574	08675
2	19787	66937	91769	13399
3	16746	77983	18061	23664
4	91039	16099	38824	00778
5	11075	62081	88977	78676

One of the ways to use this table is to assign a number to each member of the population. Then make two arbitrary decisions:

1. Decide to read the table either vertically or horizontally.
2. Select a starting point.

---

*Adapted from Walker, Helen, M. and Levy, J. Statistical inference. N.Y.: Holt, Rinehart and Winston, 1953.

The starting point can be anywhere (upper left-hand corner, third column line 4, lower right-hand corner, etc.). Suppose, for purposes of illustration, you decide to start in the upper left-hand corner and read vertically. Suppose also that the population in question has 350 members in it. You want to select 75 persons in the sample. The task is to locate the first 75 numbers that fall in the range of 1 to 350. Starting with the first number, 50691, look at the first 3 digits. 506 is not in this range. Go to the next number. Since 197 is in the range, student 197 is the first to be selected in the sample.

Check Your Random Sample before Collecting Data

If your sample is small (e.g., classroom units), it is a good practice to check the distribution of important group characteristics before collecting data. By chance, a sample may be drawn which over- or under-represents some variable you wish to study. For example, you may draw a classroom with 20 boys and 5 girls, or you may draw a sample without the ethnic representation you wish to have. Some samplers advocate that a second or third independent random sample be drawn if this should happen.

LEARNING EXERCISE 8: RANDOM SAMPLING

Complete the following exercise for randomly selecting five (5) students for classroom observation.

a. Number each name in the list of students below.
(Start with LEFT column.)

______	Paul Adler	______	Tom O'Toole
______	John Allen	______	Brian Peters
______	Mary Brummer	______	Andrew Ramirez
______	Ken Duman	______	Margaret Smith
______	June Feng	______	Sheri Thompson
______	Scott Goldsmith	______	Carmen Thurber
______	Ann Jamison	______	Terry Ting
______	Yoko Kimoto	______	Phyllis Unwin
______	Cathy Labovitz	______	Rodney Woods
______	Jerry Mann	______	Roy York
______	Carolyn Mendez		
______	Ramon Nunez		
______	Pat O'Conner		

b. Check (✓) which way you will read the Table of Random Numbers on page C-30:
Horizontally (across) ______ or vertically (up and down) ______.

c. Check (✓) where you will start reading the Table of Random Numbers:
Left upper corner ______ or right bottom corner ______.

d. Identify the number of digits necessary for selecting the sample from the list of students given above: ____________.

e. Using the Table of Random Numbers on page C-30, follow the instructions given below:

1. Read the first two digits in each five-digit number.
2. Use the Table of Random Numbers and identify the names of the students in the list on page C-28 that correspond to the numbers in the table.
3. Using the method you choose to select the random numbers (see b and c on page C-28), enter the students' names in the appropriate column below.

Students for Classroom Observations

Students To Be Observed	Method 1: Horizontal Left Upper _____	Method 2: Horizontal Right Bottom ____	Method 3: Vertical Left Upper _____	Method 4: Vertical Right Bottom ____
1.				
2.				
3.				
4.				
5.				

TABLE OF RANDOM NUMBERS (DIGITS)

Row No.										
1	50691	91653	88574	08675	12700	32027	41034	56912	34264	77769
2	19787	66937	91769	13399	96096	43165	72096	86350	23062	99419
3	16746	77983	10861	23664	64557	78213	43857	68009	20483	00618
4	91039	16099	38824	00778	23058	76539	50584	71810	52589	32778
5	11075	62081	88977	78676	53855	56472	13090	01708	89016	45111
6	41230	92934	30342	29933	24597	72632	21727	63861	80454	47243
7	59028	24399	05075	64775	59803	45737	19025	46696	18914	03062
8	42957	25204	00753	60284	85483	34984	86637	95354	80698	98750
9	45881	59475	64445	98261	55252	50788	31295	16437	49497	22493
10	75104	45819	88471	75440	55309	63481	23616	64950	73291	10964
11	78614	07347	63528	84643	10455	95596	38158	75758	65628	10498
12	69278	59274	67459	53563	98241	18097	65297	49803	99145	25320
13	58626	91259	13832	75095	08333	53845	74223	82690	89320	89565
14	81630	00339	07996	65249	66792	05555	79169	12136	44621	95904
15	74330	13688	02044	65910	96007	82692	40473	56437	35671	95072
16	70829	66963	86390	26458	02385	41505	06239	68990	32915	89542
17	55084	58581	60759	20627	86682	76542	03648	38183	29823	68134
18	98845	17428	97397	62400	51284	92211	40593	82713	06067	46190
19	48116	91870	16346	97406	54649	42039	58407	84248	45780	60547
20	82778	31709	71564	26258	07522	03825	92087	21809	25678	39987
21	86615	67618	07446	63129	07111	70516	67289	09457	48995	08043
22	82558	99260	69136	35099	68187	85382	09569	94211	57824	98100
23	08290	70291	74090	96503	56140	27794	27765	51740	07712	29816
24	95062	76310	81603	86828	68370	46001	79205	35511	91239	52961
25	30361	66712	86801	29556	91232	98295	87322	99172	50009	27224
26	17390	96107	70391	78715	61943	33315	39778	97149	08122	86388
27	05390	33046	63920	28733	42644	38972	98161	79861	88282	28279
28	06624	21114	33869	20940	03732	39973	89948	81060	36381	06027
29	38146	77295	33742	00135	26587	54775	94846	18587	39327	71711
30	76430	28645	62335	60393	71813	52677	09917	89100	93855	75617
31	16664	30164	22546	63538	79376	26865	61995	60418	37777	84170
32	56424	64680	81038	79364	23815	44002	38380	09864	35950	10760
33	95954	15540	18554	63349	70259	03212	91950	16214	80378	56421
34	59007	56364	49965	61970	32493	55404	85950	99606	46328	17887
35	19341	87208	99853	40202	08553	78731	83463	19524	82512	13556
36	24505	87007	35748	54865	40209	49466	94574	31406	64422	87185
37	15086	92183	84632	36790	59608	00371	67456	55361	80669	75402
38	65664	02188	09164	70939	25856	24344	58859	10454	19212	59078
39	40397	76835	14062	96067	70645	23695	59140	75812	18804	55529
40	31700	24753	22919	43207	83387	27820	12494	30041	88927	22668
41	14472	19372	23759	47116	81647	44946	97716	41157	30913	30842
42	18018	57089	98428	89075	77511	15194	69634	68269	52292	63404
43	16752	54266	76103	05268	41145	36100	73916	32462	01658	68565
44	47184	33660	96555	56656	18238	56888	29315	99813	47831	81395
45	93884	63945	06606	45545	29237	21040	43552	02749	19963	23705

ANSWERS

Students To Be Observed	Method 1	Method 2	Method 3	Method 4
1	Y. Kimoto	R. York	C. Thurber	R. York
2	R. Nunez	C. Thurber	A. Ramirez	R. Woods
3	C. Thurber	J. Allen	C. Mendez	P. O'Conner
4	P. O'Conner	P. Unwin	Y. Kimoto	M. Smith
5	R. York	S. Goldsmith	M. Smith	J. Mann

Stratified Random Sampling

Another way to handle the problem is to draw a stratified random sample in which the population is first divided into categories or strata and then random samples are selected for each category or stratum. The more of these categories you include, the less you have to depend on randomization to handle the extraneous or uncontrolled factors, for the units within a sampled stratum will be alike on the category selected for stratifying.

Here are two examples:

A. Here is a population of 7th, 8th, and 9th grade boys and girls given one or two periods of reading instruction per day.

POPULATION

B. We might begin stratifying the population by choosing the factor of SEX.

BOYS
GIRLS

C. We might also choose the factor of AMOUNT OF READING INSTRUCTION. Divide the population again into the two levels of amount of reading instruction--one period and two periods per day.

	ONE PERIOD	TWO PERIODS
BOYS	Boys with one reading period per day	Boys with two reading periods per day
GIRLS	Girls with one reading period per day	Girls with two reading periods per day

D. Finally, we might want each grade adequately represented. Divide the population again into three levels by grade--7th, 8th, 9th.

	GRADE	7th		8th		9th	
	AMOUNT OF READING	One Period	Two Periods	One Period	Two Periods	One Period	Two Periods
SEX	BOYS	7th grade boys with one period of reading per day	7th grade boys with two periods of reading per day	8th grade boys with one period of reading per day	8th grade boys with two periods of reading per day	9th grade boys with one period of reading per day	9th grade boys with two periods of reading per day
	GIRLS	7th grade girls with one period of reading per day	7th grade girls with two periods of reading per day	8th grade girls with one period of reading per day	8th grade girls with two periods of reading per day	9th grade girls with one period of reading per day	9th grade girls with two periods of reading per day

The strata or subgroups from which we are sampling are clearly becoming more and more homogeneous. Our originally fairly heterogeneous population with its characteristics of 7th, 8th, and 9th grade boys and girls with one or two periods of reading per day has become 12 smaller, more homogeneous subpopulations. Random samples from these smaller, relatively more homogeneous groups yield more representative samples, although fewer students are drawn into them.

A. If you randomly sample 48 items from a population of test items which contains within it six subtests,

SAMPLE 48
ITEMS FROM
ENTIRE TEST

then the random sample may include more items from one subtest than another by chance.

B. However, if you divide the population of items into subtests first and randomly sample items within subtests equally,

SAMPLE 6 ITEMS FROM SUBTEST 1	SAMPLE 6 ITEMS FROM SUBTEST 2	SAMPLE 6 ITEMS FROM SUBTEST 3
SAMPLE 6 ITEMS FROM SUBTEST 4	SAMPLE 6 ITEMS FROM SUBTEST 5	SAMPLE 6 ITEMS FROM SUBTEST 6

then your sample includes an equal number of items from each subtest and you can discuss the results more definitively in terms of subtests and the test as a whole. You could also select a variable number of items per subtest depending upon where you want to put the emphasis or according to the proportionate allocation of items in the original test. The important thing is that you have more control over composition of the final sample.

Matrix and Multistage Sampling

Another kind of sampling closely related to the stratified random type is *matrix sampling*. In this instance, the instruments or items are sampled. If data are needed on a large number of objectives, for example, rather than subjecting one sample of students to a lengthy test or series of tests, the evaluator administers samples of the test items or tests to different samples of the population.

Multistage or cluster sampling is a technique of random sampling that is frequently used. The most used method in surveys is the *successive* random sampling of units (or groups and subgroups). For example, in a statewide evaluation, the evaluator first randomly selects districts, then schools within districts, then classrooms within schools. This amounts to a narrowing down, in stages, of the sample with a randomness procedure at each stage.

9. HOW LARGE SHOULD A SAMPLE BE?

Sampling within a school or district for program-evaluation purposes is not practical except for the larger schools and districts. However, sampling of parent or community groups is practical for all except the smallest of communities. The size of the population and the amount of error the evaluator is willing to tolerate is what determines the practicality of using a sample. "Population" in this sense means the group for whom you want information--it may be all fifth graders or all parents of secondary school students enrolled in noncollege preparatory curricula or all adults in the community of voting age.

Whenever a sample is used, the inevitable question which must be faced is: What would the results have been if everyone in the population had been included? The sample, if appropriately drawn, gives an estimate of what the results would have been for everyone in the population had it been possible to include everyone. However, there is always presumed to be some error in this estimate. The evaluator does not know how much error there is but he or she can control the expected amount of error by selecting a sample proportionate to the number of persons in the total population. Suppose, for example, you ask a sample of parents if they approve some organizational change the board of education is considering. The change is a fairly major one, and you want to be sure that the proportion of sampled parents who approve comes within 5 percentage points (with 90 percent certainty*) of what the results would have been if all parents had been asked. If there are 2,000 parents, you would need to obtain 322 responses to achieve this degree of accuracy. The table that follows was developed using this degree of accuracy and shows sample sizes for various population sizes.

---

*Sampling cannot guarantee certain variation 100 percent of the time, but it is possible to know how sure you can be of your results.

TABLE FOR DETERMINING SAMPLE SIZE FROM A GIVEN POPULATION*

N	S	N	S	N	S
10	10	220	140	1200	291
15	14	230	144	1300	297
20	19	240	148	1400	302
25	24	250	152	1500	306
30	28	260	155	1600	310
35	32	270	159	1700	313
40	36	280	162	1800	317
45	40	290	165	1900	320
50	44	300	169	2000	322
55	48	320	175	2200	327
60	52	340	181	2400	331
65	56	360	186	2600	335
70	59	380	191	2800	338
75	63	400	196	3000	341
80	66	420	201	3500	346
85	70	440	205	4000	351
90	73	460	210	4500	354
95	76	480	214	5000	257
100	80	500	217	6000	361
110	86	550	226	7000	364
120	92	600	234	8000	367
130	97	650	242	9000	368
140	103	700	248	10000	370
150	108	750	254	15000	375
160	113	800	260	20000	377
170	118	850	265	30000	379
180	123	900	269	40000	380
190	127	950	274	50000	381
200	132	1000	278	75000	382
210	136	1100	285	1000000	384

NOTE: N is population size.
S is sample size.

---

*Krejcie, Robert V. and Margan, Daryle W. Determining sample size for research activities. Educational and Psychological Measurement, Vol. 30, 1970, 607-610.

Clearly, sampling within classrooms is not appropriate for program evaluation purposes. However, sampling on small populations (such as a classroom) may be used for other purposes. Exploratory or pilot studies may give indications or hunches which can then be studied more thoroughly with the larger groups. Groups of between 10 and 30 can be used advantageously for such purposes and are easier to handle computationally.

10. TWO CASE STUDIES ON SAMPLING

Number 1

The English teachers at a high school had been receiving considerable criticism from some members of the community whose sons and daughters had not performed well on the Scholastic Achievement Tests. The English teachers did not want their program judged on this single criterion and therefore wanted to gather and publish information about achievement on the full spectrum of objectives in the English program. They developed a plan and identified and developed a series of instruments to assess objectives in five different areas:

- Reading -- A standardized test will be used with scoring by the publisher.
- Spelling -- Teachers will read a list of commonly misspelled words to students who will write the words on a piece of paper.
- Familiarity with literature -- A multiple-choice test developed by the English department faculty will be given by the teachers.
- Appreciation of literature -- Students will be interviewed by someone other than their own teacher.
- Writing -- Teachers will evaluate a paragraph written by each student.

Having done this planning, the teachers realized that the evaluation process would consume enormous amounts of teacher and student time and would cost considerable money unless some economies were possible. They decided to explore the possibilities of using sampling procedures.

As an educator recently sensitized to some of the advantages and disadvantages of sampling procedures, what advice would you offer? For which objectives would you sample and for which would you use the entire student body? Why?

The following suggestions may be similar to some of your considerations:

- Reading -- The costs of administering and scoring these tests are not great. The information may be needed for student educational guidance and placement anyway. Therefore, it is probably appropriate to test all students at least at entrance and before they leave school. Students with deficiencies might be tested at intervals. Moreover, all parents are probably concerned about their own children's achievement on this measure.
- Spelling -- The situation is analogous to the reading objective, and the same recommendation would seem appropriate.
- Familiarity with literature -- This test is fairly easy to administer and score. However, it may not be equally important for all students. Nor will all students receive equal exposure. Sampling of different groups of students based on courses taken or tentative plans after high school would seem sensible.
- Appreciation of literature -- Since an interview is planned, the time required will be considerable. Training will be necessary if the interviewer is to avoid biasing the responses of the person

interviewed. Since the interviewer is not to be the student's own teacher, arranging for someone to do the interviewing may be difficult or costly. Sampling would seem very appropriate.

- Writing -- Sampling would provide convenience. Students are undoubtedly writing paragraphs for some of their regular assignments. These regular assignments could be the source of the sample, provided the teachers could agree on a common assignment for this purpose. To avoid bias, each teacher might read and criticize the first paragraph of a regular assignment turned in to another teacher. In this case, all students might be assessed with sampling to avoid a great amount of testing and evaluating concentrated at one or two times of the year. Sampling of regular assignments might also encourage consistency in quality of writing as opposed to a one-time effort.

Sampling is probably used less frequently than it could be for program evaluation. It should be remembered that sampling procedures can be applied not only to students but also to test items, time of day, teacher performance, textbook content, and so on. Sampling fits any of the many larger populations about which we may want information. We can often get as much information as we need by applying our measurement to a relatively small part of the whole.

Number 2

The preceding discussions have emphasized the use of samples and especially of random samples. The election of 1936 provided one of the classic cases of nonrandom sampling and its consequences. The Literary Digest magazine had correctly predicted several preceding elections and used its tried and tested technique in 1935 when ballots were mailed and 2,300,000 returns were received. Based on analysis of these data, the Literary Digest assured

Republican Alfred Landon that he would defeat Franklin Roosevelt by 241 electoral votes to 99. The election did not work out that way, and the case illustrates two important points:

1. The Literary Digest used a sample of 2,300,000. In recent years, Gallup polls have used 2,000--4,000 but have obtained more accurate results.
2. The Literary Digest drew its sample from lists of people such as telephone directories. But at that time, not all people had telephones. In particular, low-income people were less likely to have telephones, especially during the depression. In the nineteen-twenties, income level had not been a significant predictor of political preference as it was in 1936. The Digest had drawn conclusions about the voting population from a sample which was not representative of the population. Had the sample of voters been random, each income group would have shown up in the sample proportionately to its share of the population. The same would have been true for geographic regions, race, age, and any number of other factors that might have influenced voting patterns.

11. A FINAL WORD ON DESIGN AND SAMPLING

It is clear that good evaluation design produces information that is valuable to schools. It is also clear that the cost of translating such design into a working evaluation of a program is frequently high. Mistakes must be kept to a minimum. Thus, it is essential to approach the design of an evaluation from the pragmatic as well as the theoretical point of view.

Here are some practical considerations to keep in mind when approaching any program evaluation design:

1. You need to keep records. In order to compare present programs with past programs and to gauge progress, you must have adequate information. You may need the cooperation of other schools.
2. Use the same or comparable instruments in different times and places. If you keep changing instruments, you can never make comparisons over time. Fortunately, sampling procedures make it possible to continue using the same instruments while also adopting new ones.
3. Resist the natural impulse to treat all students alike if you want to assess the effects of different programs.
4. You need to be able to hold programs still long enough to look at them. This means that innovation cannot be constant, but should progress in planned and measured increments of improvement and change.
5. You need to plan much farther in advance and include planning for evaluation as part of program planning.
6. You need to communicate clearly to students, parents, and teachers why you are doing what you are doing.

In this section, we have also indicated ways in which evaluation designs can be applied more economically through the use of sampling techniques. Sampling is particularly suited to program evaluation which is based on information about groups, not individuals. Sampling exposes only a portion of the program population to evaluation procedures; if done properly, however, the information produced on a relatively small segment of the population will be comparable to what might have been produced on the entire population.

PROGRAM EVALUATOR'S GUIDE

Section D

SELECT OR DEVELOP ASSESSMENT INSTRUMENTS

The Evaluation Improvement Program

PRECIS

For the most part, the program evaluator will prefer to select assessment instruments from those already available, though under some circumstances, local development of specially devised instruments will be more appropriate. Many types of instruments are available; which type to select depends, of course, on the objectives of the program.

An instrument should be selected for each discrete learning outcome that is expected to result from the program. Skills, abilities, knowledge, and understanding are outcomes best measured by tests. Attitudes, feelings, and appreciations are more appropriately measured by questionnaires and structured interviews. Behaviors, interactions, and practices may be more satisfactorily assessed by means of observation instruments. High-priority objectives will require multiple measures.

Careful selection and development of instruments for program evaluation help assure that all the information needed to judge how effectively objectives are met will be available when the data-collection effort is complete.

CONTENTS

Page

1. INTRODUCTION D-1
    Scenario 1 . . . . D-2
    Scenario 2 . . . . D-3
    Scenario 3 . . . . D-6

2. OVERVIEW OF THE EVALUATION MODEL D-11

3. CONSIDERATIONS IN SELECTING ASSESSMENT INSTRUMENTS D-11
    Standardization D-13
    Reliability and Validity D-14
  LEARNING EXERCISE 9: RELIABILITY AND VALIDITY D-17

4. TYPES OF ASSESSMENT INSTRUMENTS D-20
    Achievement Tests D-21
    Questionnaires D-27
  LEARNING EXERCISE 10: JUDGING ITEMS D-43
    Observational Techniques and Instruments D-48
  LEARNING EXERCISE 11: CRITICIZING A CLASSROOM OBSERVATION INSTRUMENT D-59
    Other Behaviors D-63

5. SOURCES OF INFORMATION ABOUT INSTRUMENTS D-66
    Use Multiple Measures Whenever Possible D-68
  LEARNING EXERCISE 12: SELECTING NORM-REFERENCED TESTS D-69
  LEARNING EXERCISE 13: SELECTING APPROPRIATE INSTRUMENTS D-74

6. LOCATING EXISTING INSTRUMENTS vs. DEVELOPING ASSESSMENT INSTRUMENTS LOCALLY D-76
    Developing Instruments D-76

7. REVIEW D-79

1. INTRODUCTION

Now that you have seen the kinds of questions that must be addressed in the planning stages of a program evaluation, we are going to take time, through a set of three scenarios, to show what more typically happens in school districts.

List of Characters (in order of appearance)	Personality Type
1. Mrs. Smith	Classroom teacher, disgruntled with evaluation report on her classroom
2. Principal	Sympathetic to Mrs. Smith but painfully trying to meet state reporting requirements
3. Chairman of the Board of Education	Responsible member of the community trying to look out for the school's and the taxpayer's interest
4. Mr. Worth	The evaluator, who is trying to do the best job he can within the constraints imposed
5. Parent 1	Mother of two children in school--for the program
6. Parent 2	Father of children in school--but wants to know he's getting his money's worth
7. Boy	An eighth grader who thinks the program is "far out"
8. Mr. Fairchild	A district supervisor of evaluation who audits evaluation plans

Scenario 1: The Report: Useful to Whom and for What?

Setting:	(A principal's office. You are the evaluator of a reading program waiting to see the principal of the school who is busy talking with Mrs. Smith. You know Mrs. Smith. She is a reading teacher and you have been in her classroom once or twice. Your impression is that she is a very outspoken type of person and that she often talks a great deal. The door of the principal's office is not closed, and you hear the following conversation.)
Mrs. Smith:	Did you read that dumb evaluation report? That evaluator doesn't know what he's talking about. He's only been in my classroom twice during the whole year and then only for twenty minutes! Yet he concluded NO SIGNIFICANT GAINS. I'm a teacher. I don't know about all this evaluation stuff. All I know is my kids and the terrific progress that some of them are making. The whole class has improved in reading! They enjoy reading in a way you wouldn't believe! I'm proud of them, and I certainly won't let my kids be put down by some fancy evaluator.
Principal:	Mrs. Smith, we all know that evaluation reports don't show what's really going on in the classroom. They're not supposed to. These reports are only for the central office and the capital. They make us send a report. I don't think anybody takes the time to read them. They certainly don't affect my feelings about our program.
Mrs. Smith:	That may be so. But I think people in the capital have the right to know about the good things that are happening here. You don't need to be a professional evaluator to figure out by yourself that all the students are different, that they learn and progress at different rates and that you ought to teach accordingly.

Mrs. Smith: (cont'd)

He used just one test, looked at some one thing that he called the MEAN and decided that the program was no good. That evaluator certainly is MEAN, not to mention unfair and overpaid. Many of these tests aren't even related to what we're teaching in our reading program. It's obvious that the individualized instruction program that we started with such a big effort is helping students. Even the parents noticed the improvement in their kids. We had an individualized program just like we talked about in those workshops. Incidentally, those were really good workshops and there isn't a word in that evaluation report about those either.

Scenario 2: The Program: Payoff or Ripoff?

Setting:

(A school auditorium. You are in the audience along with many interested parents at a meeting of the Board of Education. On the stage, seated at a table, are members of the board, the principal, the program planner/evaluator, and a teacher who participated in the program. The main item on the agenda is the experimental [demonstration] reading program. The Board is meeting to collect facts pertaining to the impact of the program. On the basis of this information, the decision will be made as to whether or not to continue the program. The meeting has been under way for a short time.)

Chairman:

. . . so we've called this meeting as part of our responsibility to the community to see that its school tax dollar is getting the best return for the investment.

I have the report that Mr. Worth, the evaluator, has prepared. Mr. Worth, let me start by asking you a broad, general question. Given the fact that you find "no significant gain," do you feel that there is any justification for carrying the program for a second year?

Mr. Worth: That's not a simple question to answer. You're quite right in pointing out that we found no significant gain in reading achievement. However, on page 37 of the report, I presented my criticism of the evaluation. Let me repeat now, for the sake of those who may not have seen the report, that my total evaluation budget was $2,000. Furthermore, no consideration was given to an evaluation until school opened. In this situation, we tried simply to meet the minimum district and state requirements. What we did was administer a pretest in September and another form of the same test in June. It was a standardized reading achievement test and it didn't really reflect all parts of the experimental program. If I had it to do over, properly, you can be sure that this evaluation would look quite different. For example, you'd have data on how the kids feel about the program.

Mrs. Smith: Mr. Chairman, I was quite disappointed in the evaluation report. I was telling Mr. Jackson, our principal, not too long ago how misleading I thought it was. Do you now that some students were so excited about the program that they'd come early in the morning and often stay after school just to work out extra assignments?

Chairman: Thank you, Mrs Smith. We made this an open meeting because we felt that interested members of the community ought to be heard. Does anyone in the audience wish to ask a question or make a comment? Please speak up and start by stating your name. . . Yes, madam.

Parent 1: My name is Mary Thatcher. Two of my children are in Gardenview Elementary School. Tommy's going into grade 7 this year. We've heard a lot about the new reading program. My husband and I know several parents with children in the

Parent 1: cont'd — program. They're extremely pleased with the progress of their children. Tommy could sure use individualized instruction--he's a good boy, but he doesn't read too well, and this program sure sounds like just what he needs.

Chairman: Thank you, Mrs. Thatcher. . . Your name, sir?

Parent 2: I'm Farley Grant. We've lived over on Oak Street, paying property taxes, for well on 10 years now. Now that we have school-age children, we want them to learn something. I know that you can't get by these days without knowing how to read, and I want my kids to get the best possible start. But I want to know that my tax money isn't being wasted on some fad or other. I learned to read without all these fancy frills.

Chairman Mr. Grant, we appreciate your views. That's why we're having this hearing. . . Yes, son, tell us your name.

Boy: My name's Jerry Bilford. I'm in eighth grade and I just want to say that the program is really far out. I mean, I used to hate reading; but now it's really got me going and I think you should let it go on.

Chairman: Well, ladies and gentlemen, I must confess that we're really in a dilemma. I should let you know that ever since word got out that the district was considering dropping this program, my office has received quite a few letters and phone calls urging the board to keep the program alive. Hearing from you today, I get the same feeling. But, frankly, there's no hard evidence to say the the program's worth the investment.

Mr. Worth, this Board owes you an apology. Your counsel about the importance of an adequate evaluation fell on deaf ears last summer. It's been a costly lesson for all concerned.

Chairman: (cont'd) The members of the Board would like some more time to digest the information we've gathered today. We'll probably continue the program. Mr. Worth, we'd like to ask you to draw up the plans for what you consider an adequate evaluation, along with a proposed budget, for consideration by the district office.

Scenario 3: Criticizing the Evaluation Design

Setting: (Mr. Worth has sought out his colleague and most respected critic, Mr. Fairchild, to discuss the task the School Board chairman has given him. Mr. Fairchild is an auditor. His role is to criticize evaluation plans to ensure that they provide adequate information about whether or not a program is meeting its objectives.)

Mr. Fairchild: Tom, I understand that the School Board chairman apologized to you in public for criticizing your evaluation of the reading program.

Mr. Worth: Yes he did, Steve, and I don't mind telling you that I felt relieved to have him off my back.

Mr. Fairchild: I don't blame you. After all, if the man knew something about statistics, he could really have embarrassed you (laughing good-naturedly). Tom, confess now, you didn't really give this evaluation much thought, did you?

Mr. Worth: Just between you and me, I was so damned mad at their attitude toward evaluation, I wasn't really enthusiastic. They don't look at the data anyway, they just do what's popular.

Mr. Fairchild: Could you give some examples of what you would have done differently, even within the limits which were imposed?

Mr. Worth: Well, for starters, I could have argued more forcefully against the need to test every student. With the money saved on the cost of standardized tests, I could have afforded to administer a criterion-referenced test, to have interviewed some students, and to have conducted a more sophisticated data analysis. I also would have fought to delay pretesting by one week. That would have given us enough time to plan a stratified random sample of students. You know, Mrs. Smith was right. The kids who really needed individualization did benefit from the program. If we could have selected two samples of students--those identified by their teacher as those who needed individualization and those who didn't--I bet we could have shown significant gains for the first group.

Mr. Fairchild: Maybe so. In that case, you would have evidence of a relationship between student characteristics and special treatment--a really interesting and useful finding. But let's turn to your present task. What kind of proposal are you going to make for this year's evaluation?

Mr. Worth: It's going to be based on four points:

1. Clear specifications of program objectives in terms of achievement gains, attitude changes, and changes in incidental behaviors on the part of the participants

2. Multiple measures for each objective, using a wide range of instruments

Mr. Worth: (cont'd)

3. Stratified random sampling, and inclusion of a variety of student factors, instructional factors, and environmental factors in the evaluation design

4. Use of sensitive statistical tests of significance

Mr. Fairchild: Well, Tom, you've shown me once again what a skilled job of conceptualizing you're capable of. But you're an evaluator not a statistician; you'd be the first to admit that. How do you propose to handle the data analysis?

Mr. Worth: You're right, Steve. One of the things I want to budget for is the service of a competent statistician so we can get the most out of the data.

Mr. Fairchild: Let's do some reality testing now. Are you really going to be able to carry out this plan with your limited time and limited staff? Perhaps you'd better give some thought to priorities and prepare some alternatives in case you're not given more released time to devote to the evalution.

Mr. Worth: Yes, that makes good sense. I'll work up several possible plans--what I consider the ideal evaluation--and two or three alternatives which meet minimal criteria, and a cost estimate for each.

Mr. Fairchild: Perhaps you should also prepare and submit a list of those parts of the evaluation you consider essential to the program planner. Then you and he can meet and come to an agreement on an evaluation plan that addresses his needs as well as yours.

Points Illustrated in the Scenarios

- To be thorough and to be credible, evaluation should encompass the processes as well as the outcomes of a program.
- Instruments used should be matched carefully to specific program objectives.
- More than one instrument should be used for each program objective.
- Evaluation should include as many of the program objectives as possible, not just one.
- The people in the program should be consulted and involved in planning and conducting the evaluation.
- If evaluation is to be respected, it must provide information useful to people in the program.
- Adequate time and money should be provided for the type and amount of evaluation needed.
- Multiple instruments should be used to measure all feasible program goals.
- Technical procedures should be used to facilitate more economical and more thorough evaluation.
- All evaluators need outside assistance occasionally and should seek it when appropriate.
- Evaluation plans should be designed in accordance with the time and resources available.

Summarizing the Scenarios

We've followed Mr. Worth as he overheard a conversation between the principal and Mrs. Smith who was complaining about the quality of the evaluation.

We've attended a meeting of the School Board in which it was made clear that while the community, teachers, and School Board are convinced that the program has some merit, it was evident that the evaluation report did not support that proposition.

And we've overheard a conversation between Mr. Worth and his colleague, Mr. Fairchild, in which Mr. Worth has sketched an evaluation plan which will provide the kind of information that was absent in the previous year's evaluation.

We have used these scenarios for two major purposes:

- Our discussion has established some general principles about evaluation.
- The reading program discussed in the scenarios is the setting for more discussions of evaluation skills in the subsequent parts of this *Guide*.

2. OVERVIEW OF THE EVALUATION MODEL

To place the selection of instruments in proper perspective, it may help to review again the elements that go into program evaluation.

Elements in Program Evaluation

- Purpose and Requirements
- Plan and Procedures
- Evaluation Design
- Assessment Instruments
- Data Collection
- Data Analysis
- Preparation and Interpretation of Reports
- Application of Findings

This section will present some basic considerations in the selection of evaluation instruments, a review of a wide variety of instrument types, a discussion on sources of information about instruments, and an outline of important steps which must be taken if instruments need to be developed locally.

3. CONSIDERATIONS IN SELECTING ASSESSMENT INSTRUMENTS

Identifying the appropriate evaluation instruments for measuring pupil attainment of a program objective is one of the prime tasks involved in the preparation of a useful evaluation plan. It is also one of the most difficult. The chief criterion for selecting appropriate instruments is whether or not they can adequately measure the outcomes specified by the performance objectives.

Important questions to consider in identifying assessment instruments are as follows:

Does the instrument measure what it is supposed to measure? This question refers to the validity of the assessment instrument. Three kinds of validity are important to consider. First is content validity, which assesses whether the test measures the content of the program being evaluated. Second is concurrent validity which compares the test scores with other similar measures. Third is predictive validity, which tells how well the score can be used to predict future performance. A fourth and more difficult kind of validity is construct validity, which refers to the psychological processes revealed by the pupil's behavior during the test. For example, comprehension skills measured on certain reading tests are thought to evaluate a child's ability to make inferences. Evidence should be offered by the test publisher that questions on the test actually do measure this ability.

If the instrument is administered more than once to similar groups, or the same group, will it yield consistent results? This question refers to the reliability of the assessment instrument. When choosing a test, the user will want it to be a reliable measure of how much a pupil knows and how well he is able to apply his skills. The test results should be earned and under no circumstances arrived at by luck, guessing, or other chance factors. The test should be constructed so that one has confidence that the score the pupil receives will be similar to the score he would receive if the test were administered to the same person again.

Is the instrument appropriate for use on the population to be assessed? This question refers to the following:

- Grade-level appropriateness
- Ethnic appropriateness
- Compatibility of norms to the groups
- Appropriateness of instructional content

Does the instrument yield objective data? If it does not, how will you control for observed differences among those collecting the data?

Is the instrument easy to administer and score? For example, interviews using structured guides are generally difficult to administer and to score, although sometimes they may be needed measures.

What time and resources are required to administer and score the instrument? As an example, individually administered instruments require more time and resources than instruments given to groups.

How disruptive is the administration of the instrument to classroom learning activities?

Will the instrument provide data which are useful for decision making at both the classroom level and the school and district level?

Is the cost of purchasing the instrument reasonable and within the allocated budget?

Each of these questions should be carefully reviewed during the process of selecting appropriate assessment instruments.

When the program evaluator considers what instruments might best measure a program's objectives, he or she needs to know the meaning of standardization and the importance of reliability and validity. The following discussion provides a brief review of these concepts.

Standardization

Standardization implies different things to different people. For the purposes of this discussion, if an evaluation instrument has the following characteristics, it will be considered standardized:

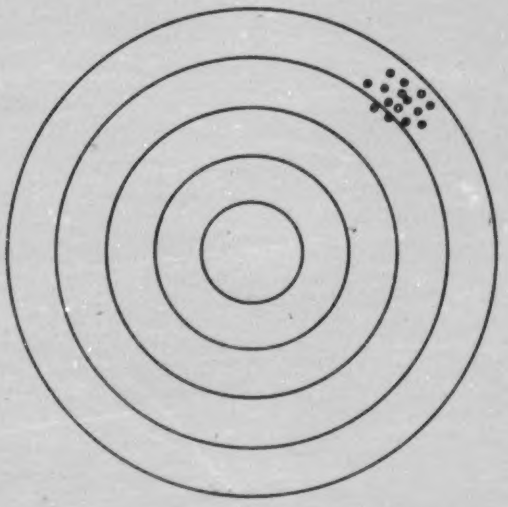
General Characteristics of Standardized Instruments

- Items are systematically structured.
- Specific directions are given on how to administer the instrument.
- Definite instructions explain how to deal with the information secured.
- Evidence is available on validity and reliability.

Mention of norms has been omitted from the list so that the broader definition of standardized instruments might apply to some criterion-referenced tests, questionnaires, and observation records.

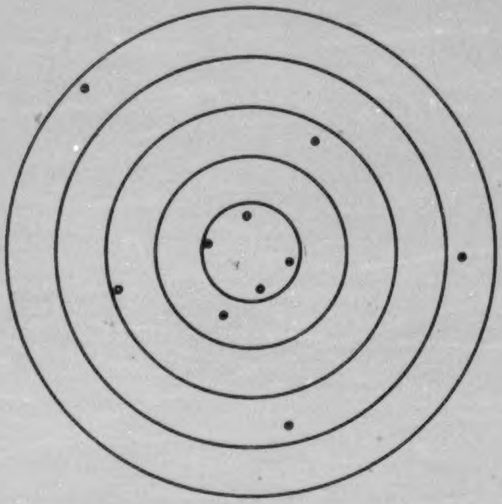
Reliability and Validity

Reliability and validity are two important characteristics of evaluation instruments. The reliability of a measure indicates the extent to which it is consistent in measuring whatever it is meant to measure. Suppose, for example, that a rifle placed in a vise were fired several times at the bull's-eye shown below, and that the bullet holes formed a tight cluster, as shown. In this case, the setting of the rifle would be reliable in that the bullets hit the same area of the target each time the rifle was fired. The validity of a measure, however, indicates the extent to which an instrument measures what it is designed to measure. In this case, the setting would not be valid because none of the bullets hit the desired target (the bull's-eye). Now suppose that all the bullets were spread all over the target, as shown on page D-15. Even

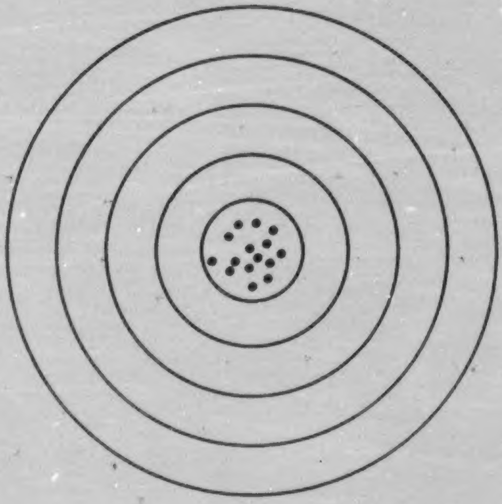


RELIABLE BUT NOT VALID

though the bullets from the vise-held rifle hit the center of the target several times, there were some stray shots, which indicates a degree of inconsistency in the way in which the rifle performed. Thus, the setting of the rifle in this case was neither reliable nor valid. The desired result would be as shown at the bottom.



NEITHER RELIABLE NOR VALID



RELIABLE AND VALID

The instrument that will be most useful in program evaluation will measure in a consistent way (reliability) what it was intended to measure (validity).

Validities of standardized tests often are expressed as validity coefficients, numbers that express a degree of relationship, generally between sets of scores from two different measurements. However, the type of validity that is most important to program evaluation is content validity. Content validity is arrived at judgmentally, by comparing each item in an instrument to the objectives of the program. The key question to ask is: Does this item measure an outcome the program sought to accomplish?

It is helpful when you do attempt to interpret reliability and validity coefficients to have some guidelines as to what is acceptable, even though there are no hard and fast rules. In general, reliability coefficients can be expected to be higher than validity coefficients, primarily because of the fact that reliability is determined either on a single instrument or between parallel forms of instruments, and validity is determined by two different assessments of the same content.

Guidelines on

Reliability and Validity Coefficients

Reliability	.80-.99	High
	.50-.80	Questionable
	Below .50	Unacceptable
Validity	Above .75	High
	.50-.75	Acceptable
	Below .50	Questionable

LEARNING EXERCISE 9: RELIABILITY AND VALIDITY

Directions: For each statement below about an instrument's characteristics, identify the explanatory statement about its validity and reliability that is most likely to be true.

Explanatory Statements:

A. The instrument is both valid and reliable.
B. The instrument is valid but not reliable.
C. The instrument is reliable but not valid.
D. The instrument is neither valid nor reliable.
E. Not enough information is provided to make one of the above decisions.

_____1. In an attempt to measure overall reading achievement, you have found that the test you are using correlates highly with a widely accepted test of social studies and moderately with a widely accepted test of reading comprehension.

_____2. A student questionnaire is administered to a group of students at two different times, three weeks apart. The two sets of scores are very similar, student for student. In addition, the items on the questionnaire have been reviewed and accepted as important and relevant by both faculty and student reviewers.

_____3. Even though an instrument you have selected seems to be measuring your instructional objectives, you find that the scores for any given student vary widely when the instrument is used the second time: An appreciable number of students do less well the second time. You are able to rule out extraneous influences such as physical environment, teacher performance, etc.

_____4. This arithmetic test you are reviewing for possible use is found to correlate very highly with the Standford-Binet.

_____5. A parent questionnaire you are planning to use is judged to have items more appropriate for teachers than for parents. In addition, on two preliminary field tests given to the same group of parents two weeks apart, you discover very little consistency on what a person does the second time as compared to the first time.

_____6. You have found an instrument to use in classroom observation that comes to you highly recommended by a friend of yours in a neighboring district. His main caution is that the results you get may be heavily dependent on just who does the observation. However, it is evident that the instrument is designed to measure those things you are more interested in observing.

_____7. You have discovered that an unobtrusive measure you have been using the last three years gives you results which are amazingly stable. However, the new goals established for the district make you think that this measure may no longer be appropriate.

ANSWERS

Because of the highly judgmental nature of some of the issues underlying some of the situations described, four research scientists at the American Institutes for Research were asked to key these items. Compare your answers with theirs.

1. No consensus. Both A and C can be defended. If a test correlates highly with some acceptable test seemingly unrelated to it, it must be reliable. Hence C would be an appropriate answer. However, in this case, the social studies test could be highly loaded verbally and could be testing reading as much as social studies. Therefore, the test could be both reliable and valid.
2. A. Consensus
3. D. Consensus If you selected B, remember you cannot have validity without reliability.
4. No consensus. Three of our research scientists said E. One said C on the same basis that number 1 could be keyed C. If the arithmetic test correlates highly with some respected test, it must be reliable.
5. D. Consensus
6. No consensus. Both D and E can be defended. If the friend's caution about results depending on who does the observation means it is impossible to get inter-rater reliability, the answer is D. If the caution implies the results depend on a high degree of experience and training in use of the instrument, it may be both reliable and valid. Not enough information is given to make this decision, hence E.
7. C. Consensus

4. TYPES OF ASSESSMENT INSTRUMENTS

The selection of assessment instruments begins with the general question of what is to be measured:

What is Being Measured?

- Achievement
- Performance
- Attitudes
- Interactions among persons
- Other behaviors

Depending upon what is to be measured, one of four kinds of instruments will probably be used:

What Kinds of Instruments Will Be Used?

- Achievement tests
- Questionnaires
- Observational records
- Logs (pupil/teacher/school records)

In planning for data analysis, it is essential that careful attention be paid to the types of items used on the various instruments and the kinds of scores various item types yield.

What Kinds of Items Are There?

- Open-ended
- Objective
  - -true/false (yes/no)
  - -multiple choice
  - -ratings
  - -checklists
- Mixed (open-ended and objective)

What Kinds of Scores Are There?

- Raw scores
- Grade equivalents
- Percentiles
- Standard scores
- Stanines
- Categories
- Rankings
- Rating scales

The sections that follow will discuss each of these types of instruments and give examples of item types and scores.

Achievement Tests

In the past few years, criterion-referenced tests have gained in popularity until today they provide the program evaluator with an alternative to the more traditional norm-referenced tests. The basic difference between the two types of test is in their design and use. A norm-referenced test is designed to place students in rank order or to compare them with other students. A criterion-referenced test is designed to tell what a student knows, understands, or can do in relation to specific objectives that are expected to be realized.

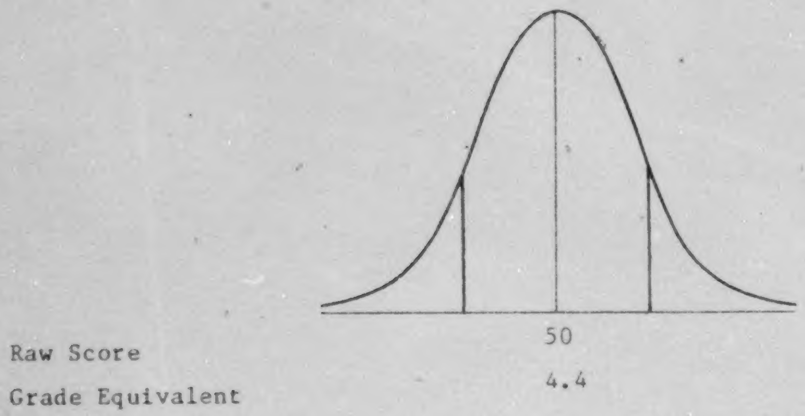
Some advocates of criterion-referenced tests say there is little need for the traditional norm-referenced test in program evaluation--that criterion-referenced tests are the only appropriate achievement tests to use. However, the question is not one of either/or. Rather, it is what kind of information you want. If you want to know how students stand in relation to some external group (other schools in the district, the state as a whole, or the nation), a norm-referenced test should be used. If you want to know where students stand with respect to some standard of mastery, a criterion-referenced test would be appropriate.

The evaluator should consider the objectives of the program carefully before deciding whether to use a norm-referenced test or a criterion-referenced test or both.

Norm-referenced tests. Funding agencies often require comparisons of the results obtained by students in the program with the general school population. If so, the use of standardized norm-referenced tests in a program evaluation is necessary.

The major disadvantage of using such a test is that it may not measure the specific content of the instruction provided in the program in question. Since norm-referenced tests are constructed to be administered to students who have been instructed in a wide range of curricula, the items cannot be expected fully to reflect the content of any particular curriculum.

When selecting a norm-referenced test, the evaluator will want to consider the kinds of scores the instrument will provide. The two most commonly used test scores are grade equivalents and percentiles. Each has its advantages and disadvantages. Grade equivalents have been particularly misunderstood and misused, both by educators and by the public. A grade-equivalent score is the mean or median score of the norm group at the time the test was normed. For example, suppose a test for fourth graders is normed in the fourth month of the school year, and on a 100-item test the average raw score is 50. A raw score of 50 is then assigned a grade equivalent of 4.4.



The actual range of raw scores may extend from 10 to 92 and the range of grade-equivalent scores assigned to different raw scores from 2.5 to 6.3. By definition, half the group is below average. It may be unrealistic to try to bring everyone up to norm unless you truly believe your lowest achieving pupils should be as good as the national average.

Criterion-referenced tests. These tests have become increasingly popular in the last six to eight years because they provide a meaningful way to measure achievement of locally set objectives. With a criterion-referenced test, an overall score is generally not obtained. Rather, a small number of test items is used to determine whether an objective has or has not been met.

There are two different kinds of performance criteria. The first, Classroom Mastery Criterion, specifies the percentage of students in a classroom who are expected to master an objective.

Classroom Mastery Criterion

70 percent of the students will be able to identify all the letters of the alphabet.

The second kind of performance criteria is Student Mastery Criterion, which refers to number of items in a criterion-referenced test that a student should be able to respond to correctly in order to show that the student has mastered the objective.

Student Mastery Criterion

To show mastery, the student should respond correctly to 60 percent of the items designed to measure a given objective.

Thus, scoring criterion-referenced tests gives percentages that relate either to the group of students who achieve at a given level or to the group of items responded to correctly.

Criterion-referenced tests, custom-made to school or district objectives, are becoming increasingly available. Five sources are listed on page 23. In addition, the ERIC Clearinghouse on Tests, Measurement, and Evaluation has published a report that cites and describes 21 criterion-referenced tests.* Commercial publishers of the more traditional norm-referenced tests are now taking steps to meet growing demands for criterion-referenced tests.

---

* ED 099 427. Knapp, J. A collection of criterion-referenced tests. TM Report No. 31, 1974.

Types of Test Scores

	Definition	Advantages	Disadvantages
Raw Score	Number of right answers obtained by an individual	Easily obtained by counting right answers, appropriate for use with inferential statistical tests	Must be changed to some type of derived score in order to make comparisons with a norm group
Grade Equivalent	A score derived from a raw score that expresses grade level as an average (e.g., 8.2 is the achievement level expected of the average student in the second month of the eighth grade.	Reasonably sound "inherent" meaning in lower grades. Uses familiar units.	Easily confused with standards. By definition, half the group it was developed on are above the average and half are below average. Difficult to compare results of different tests. Not meaningful at upper grade levels.
Percentile	The score below which a given percent of the cases lie	Widely used and easily understood. Probably best all-around type of score especially when used with percentile bands that account for the probable error of measurement.	Units along scale not equal in size. Differences near median are over-emphasized. Raw score differences between 90th and 99th percentile are much greater than raw score differences between 50th and 59th percentile.
Score	A scaled score based on the mean and standard deviation which define the distribution of scores	Has equal units through entire range of values. Has normal distribution by desing. Appropriate for use with inferential statistical test.	Not commonly used in local school settings except in large-scale national testing programs such as those provided by Educational Testing Service and the American College Testing Program. Difficult for most people to understand.
Stanine	Score % of Scores 1 4 2 7 3 12 4 17 5 20 6 17 7 12 8 7 9 4 A scaled score with a mean of 5 and a standard deviation of 2	Gives maximum information for a 9-unit scale. Reasonably easy to understand. Minimizes non-significant differences as do percentile bands.	A single unit of change is very large and so will not reflect small differences in achievement. Not widely used.

Where to Find Existing Criterion-Referenced Instruments

Where	What
1. Instructional Objectives Exchange (IOX) Box 24095 Los Angeles, CA 90024	Reading, Language Arts, Mathematics, Social Studies (K-12)
2. SCORE Westinghouse Learning Corp P.O. Box 30 Iowa City, IA 52240	Reading/Language Arts, Mathematics, Science, Social Studies(K-8)
3. Comprehenive Achievement Monitoring (CAM) Sequoia Union High School District 480 James Avenue Redwood City, CA 94063	Mathematics, Science, Geography, Business, Homemaking, Language Arts, Literature Comprehension, Foreign Languages (8- or 9-12)
4. National Assessment of Educational Progress (NAEP) 300 Lincoln Tower 1860 Lincoln Avenue Denver, CO 80203	Art, Career and Occupational Development, Citizenship, Literature, Mathematics, Music, Reading, Science, Social Studies, Writing (ages 9, 13, 17, adult)
5. ORBIT CTB/McGraw-Hill Del Monte Research Park Monterey, CA 93940	Mathematics, Reading and Communications Skills (K-12)

Both NAEP and CAM are publically supported projects, and materials are either free or relatively inexpensive.

The types of items used with most achievement tests are objective -- that is, they are multiple choice, true/false, or matching and can be scored by machine.

Questionnaires

Evaluators frequently use a questionnaire to assess opinions or attitudes of participants in a program and/or those who are in some other way associated with a program.

While there may be appropriate standardized questionnaires, most evaluators either develop or adapt items from existing, nonstandardized instruments according to their appropriateness for measuring a given program objective.

The development of even very simple questionnaires is a more exacting and demanding task than is sometimes realized. Every instrument so developed must include review and field-test steps in order to avoid ambiguous questions which may yield meaningless or invalid information. To develop good questionnaires takes talent, time, patience, and money. For this reason, a thorough search should first be made to see if good instruments exist that will fit program-evaluation needs.

Questionnaires may be administered much like achievement tests; they may be mailed to individuals such as parents; or used by the program evaluator in a structured interview with a group. The need for individual structured interviews is dictated by the circumstances and may be filled by volunteers from the community. Mailed questionnaires are especially subject to bias. The people who typically fill out and return a mailed questionnaire may be very unlike the population in general. One way to compensate for, or detect, this bias is to follow up with telephone or door-to-door personal interviews with a sample of nonrespondents, using the questionnaire as an interview guide. The interview technique may also be called for if the target population is very young, of questionable literacy, uskilled in the use of English, or if the questions are very complex.

Major Uses of the Interview Technique

- To detect bias
- With young children
- With bilingual populations
- With low socioeconomic groups
- With complex questions

If an interview is deemed best, be certain your interviewers are trained and are able to ask the question in a neutral manner without leading the person being interviewed. They should be able to recognize a vague or ambiguous response and should probe in some neutral manner such as, "Tell me more about it," or "What do you mean?" until a clear response is obtained.

Guidelines in the Review and Selection of Questionnaires. There are a number of things to consider in the selection (or development) of a questionnaire:

1. Are the questions asking only for needed information? There is a tendency among some persons who develop questionnaires to include nonessential items just because they are interesting or because he or she had always wondered about such details. Avoid trivia.

2. Are the words simple, direct, and apt to be familiar to all respondents? Education, like other professions, has a technical, sometimes mystical,jargon. If in doubt, ask one or two noneducators to read the items for understandability.

3. Are the questions clear and specific? Items that are too general, complex, or otherwise ambiguous will not get the information desired. Words such as often, occcasionally, usually, many, any, much mean different things to different people. If used, they should be defined.

4. Are any items double-barreled? For example, the question "Do you plan to leave school and look for a job next year?" is addressing two issues. Each question should contain just one topic.

5. Are the questions loaded or leading? ("Why do you think instructional method A is so successful?" assumes everyone agrees that the method is successful.)

6. Do the questions apply to all respondents? A question directed to taxpayers of the community that asks, "Do you and your wife have school-aged children?" is based on too many assumptions.

7. Will the respondents' answers be influenced by response styles? A response style is a tendency to choose a certain response category regardless of item content. Examples of well-recognized response styles are:

   - Acquiescence
     Given a choice between "agree" or "disagree," a disproportionate number of "agree" responses will probably be obtained. Instead of "Do you agree with the new school policy on flexible scheduling?" ask:

     "The new school policy on scheduling as compared with the previous policy is
     ______ an improvement
     ______ not as good
     ______ about the same
     ______ don't know"

   - Social Desirability
     Some people tend to choose answers that they think everyone else will choose rather than those that express their own opinions. So avoid using questions that have a strong social preference for agreement or disagreement.

- Ordinal or Position Bias

  If they are given a 5-point scale such as

very good	good	fair	poor	very poor

most persons will tend to avoid the extremes. This can be prevented to some degree by defining the scale points in specific terms. For example, on a leadership scale, instead of "very good," use "exceptional leader; able to take over and pull things into shape; people enjoy going along with him/her; respected by subordinates." "Very poor" might be "completely lacking; definitely a follower; does not try to convince others what is best."

Item Types in Questionnaires. Questionnaires and interview instruments usually are structured to include a combination of two major classes of items: open-ended and objective. The open-ended item offers the respondent an opportunity to give his or her own answer. The objective item forces the respondent to make a choice between two or more alternatives.

Open-ended Items

During the second year of the Evaluation Improvement Project, a follow-up study was done with a sample of first-year workshop participants. Questions were designed to find out if the workshops really caused participants to behave any differently in their approaches to program evaluation. One of the open-ended items asked, "Are you doing anything differently in relation to program evaluation this year than last year, attributable to your participation in an EIP workshop?" Two hundred four usable statements were made in response to this question. Examples of response are shown below:

- Requiring evalution process be established prior to introducing new program

- Broader approach; increased awareness of need; improved data-collection methods
- Providing more inservice for staff and aides related to objectives and utilizing test results as comparative data related to those objectives
- When planning and writing projects, more care is taken to plan for evaluation from the beginning of the project.
- I am building evaluation into the thinking-through of all department projects.
- Better process evaluation procedures and techniques--tried to build in the evaluation design rather than superimpose it.
- Spending more time selecting testing instruments to assure valid conclusions in evaluation
- Involving more people, rather than trying to handle everything needed to be done in any program evaluation
- Better preparing of objectives; better choice of instruments; better overall picture of evaluation
- Working more with other staff members on follow up utilizing test results
- Better job of evaluation of programs; better job of communicating with parents regarding evaluation; better participation of my staff in planning

How do you reduce 204 such statements to a meaningful summary of data? The task is largely a matter of applying judgment and perseverance. One way of proceeding would consist of the following steps:

1. List all responses to the question on as many pages as necessary (in this case, it took six and one-half pages to record all the information).

2. Read over 30-40 responses to get the flavor of what is being said. Are some people saying the same thing but using slightly different words? What are the key ideas that are being stated? Do categories begin to form in your mind?

3. Try listing key categories. In the sample statements, the following were among the key categories:

   - More skilled with evaluation procedure
   - Better data collection and analysis procedures
   - More involved with evaluation
   - Working more with staff
   - Better organized
   - More effective reporting
   - Better selection of test instruments
   - Greater awareness of need

4. Go back to the beginning of the list and try to classify each statement under one of these categories.

5. If you find a statement that does not fit any category, create a new one.

6. When you finish the classification, check the categories. Are two or more categories near enough in meaning and intent that they may be combined? Are there several categories with just one or two responses for each? Should they be combined into one miscellaneous category? Categorization into more than 12 or so separate categories probably results in distinctions that are too fine.

The final results on the EIP survey were presented in this manner:

Changes in Participants' Evaluation Activities
Attributable to Attendance at EIP Workshop

	Number	Percent
More Skilled with Evaluation Procedure	64	31
Better Data Collection and Analysis	25	12
More Involved with Evaluation	23	11
Working More with Staff	21	10
More Organized	16	10
Better Reporting	12	6
Changed/Adjusted Evaluation Design and Ongoing Project	11	5
Better Selection of Test Instruments	10	5
Greater Awareness of Need	10	5
Not Applicable	6	3
More Aggressive	4	2
Used Sampling Technique	2	1
Total Number of Statements Categorized	204	
Total Number of Respondents	199	

Note that 204 statements were categorized from 199 persons who responded. Most persons gave a response that fit into just one category; a few gave responses that fit into more than one.

If there is a wide difference between the number of statements categorized and the number of persons responding, the evaluator may wish to look back to find out if a few persons are being so verbal as to bias the results.

This type of data collection yields category data, sometimes called content analysis. The data collection section of this Guide suggests statistical techniques to use in treating category data.

Advantages and Disadvantages of Open-Ended Items

	Advantages		Disadvantages
1.	Provide freedom and spontaneity in response	1.	Difficult and time-consuming to score
2.	Respondents usually like being asked for their opinions; good warm-up	2.	Many open-ended items make an instrument too time-consuming for respondent
3.	Useful for determining range of responses which is not possible with objective items	3.	Responses may be related to general verbal facility of respondent
4.	Testimonials lend color to the research report	4.	Testimonials, if not balanced by more objectives evidence, result in evaluations that has little substance

Objective Items

The objective item provides the respondent with a structured response. There are several types of structures: checklists, multiple-choice items, rating scales, and rankings.

In the checklist, the respondent is given a list of items and asked to check all that apply. For example, in a follow-up study of EIP workshop participants, one item dealt with whether or not the participant had taken steps to get others to improve skills in program evaluation. Those who said they had were then asked to check which of the following actions they had taken:

_____ Encouraged staff to attend EIP workshop

_____ Conducted evaluation workshop locally

______ Circulated EIP material for review and study

______ Circulated other materials related to program evaluation

______ Talked informally with staff about problems related to program evaluation

______ Helped colleagues with program evaluation problems

______ Other ____
specify

Items that require a "yes" or "no" response are like a checklist in that both create category data and would be analyzed in similar ways. Here are two examples:

Item from a Reading Lab Questionnaire

1. Do you feel you have developed better reading habits due to this course?

Yes	No

Item from A Self Concept Questionnaire

1. I feel left out of things is class.

Yes, Like Me	No, Not Like Me

A multiple-choice item requires the respondent to make a judgment based on a specific set of alternatives.

Example 1: Teacher Judgment

Which one of the changes listed below did you find most helpful in implementing the new reading program?

______ Improved selection of curriculum materials

______ Inservice training workshops

______ Increase in number of teacher aides

______ Grouping of students

Example 2: Quantity and Intensity Scale

For what portion of your activities as a program evaluator do you receive clear and specific directions from your supervisor?

_____ For almost all activities

_____ For most of my activities

_____ For about half

_____ For few of my activities

_____ For almost none of my activities

Example 3: Amount of Time Scale

When you are working, what is the average day like for you? How often does time seem to drag?

_____ About half the day or more

_____ About one-third of the day

_____ About one-fourth of the day

_____ About one-eighth of the day

_____ Time never seems to drag.

The advantages of the multiple-choice type of item are primarily in their ease of administration, scoring, and analysis. The greatest problems relate to their careful development, avoidance of ambiguities, and reasonableness of choices. These points are discussed more fully at the end of this section.

Rating scales assign numerical values to the various responses to an item in order to spread them. That is, a rating scale gives the rater the opportunity to present his or her opinion on a continuum of judgment. Most rating scales permit the rater a choice of three to five values. For example:

Item From a Teacher Questionnaire

1. What is your overall reaction to the effectiveness of individualized instruction?

Extremely Pleased	Somewhat Pleased	Neither Pleased Nor Displeased	Somewhat Displeased	Extremely Displeased
1	2	3	4	5

Choice of an odd number rating scale allows, the respondents to adopt a neutral position. (An even number of choices would force him or her to take a position.) Before deciding on the number of scale points, decide whether or not you want respondents to take a position.

The selection of descriptors for each rating on the scale is most important. Insofar as possible, they should mean the same thing to all expected to respond. In the above example, you might instruct the respondent as follows:

Descriptor	Means, in Comparison to All Techniques You Have Used
Extremely pleased	Among the top 10 percent of techniques you have used
Somewhat pleased	Better than most but not among the top 10 percent
Neither pleased nor displeased	About average in comparison to other techniques
Somewhat displeased	Below average but not among the worst 10 percent
Extremely displeased	Among the worst 10 percent of techniques you have used

Care must be taken with rating scales to define precisely what is wanted. The following will illustrate the point:

Another type of rating scale commonly used to measure attitudes consists of a series of statements, each of which has its own scale value. Typically, the statements are arranged in order from highly positive to highly negative. The person whose attitude is being measured is simply asked to check those statments with which he or she agrees. The score is obtained by adding the values assigned to the statements checked. An example of this type of scale is given on the following page.

In the example that shows the tallying and scoring, on page D-39, the median score of 4.02 for the group falls at item 6, "It solved some problems for me." This would ordinarily represent the tendency for the group. However, there was a

Kropp-Verner Attitude Scale for Measuring Effectiveness of Meetings *

Directions: Check (✓) below only those statements which accurately reflect your personal reaction to the Evaluation Improvement Program workshop.

To Be Completed by Trainer

Check Here — **Score**

____ 1. It was one of the most rewarding experiences I have ever had. ____

____ 2. Exactly what I wanted. ____

____ 3. I hope we can have another one in the near future. ____

____ 4. It provided the kind of experience I can apply to my own situation. ____

____ 5. It helped me personally. ____

____ 6. It solved some problems for me. ____

____ 7. I think it served its purpose. ____

____ 8. It had some merits. ____

____ 9. It was fair. ____

____ 10. It was neither very good nor very poor. ____

____ 11. I was mildly disappointed. ____

____ 12. It was not exactly what I needed. ____

____ 13. It was too general. ____

____ 14. I am not taking any new ideas away. ____

____ 15. It didn't hold my interest. ____

16. It was much too superficial. ____

____ 17. I left dissatisfied. ____

____ 18. It was very poorly planned. ____

____ 19. I didn't learn a thing. ____

____ 20. It was a complete waste of time. ____

Participant's Median Score

---

* Kropp, R.P., and Verner, C. An attitude scale technique for evaluating meetings. *Adult Education*, VII(4), Summer 1957

Scoring of Kropp-Verner Scale

Directions: Tally the items checked by participants in column (2) of the form below. In column (3) write the total number of tallies Obtain totals for column (3) and find the median score.

(1) Item	(2) Checked by Participant [Tally (i.e., 卌)]	(3) Checked by Participant [Total of Tallies]	(4) MD.
1	卌 /	6	1.13
2	卌 /	6	1.58
3	卌 卌 卌 卌 //	22	2.25
4	卌 卌 卌 卌 卌 卌 卌 卌 ////	44	2.77
5	卌 卌 卌 卌 卌 ////	29	3.40
6	卌 卌 卌 卌 卌 卌 卌 卌 /	41	4.02
7	卌 卌 卌 卌 卌 卌 卌 //	37	4.44
8	卌 卌 卌 卌	20	4.96
9	卌 /	6	5.30
10	/	1	6.02
11	//	2	6.78
12	/	1	6.97
13	//	2	7.19
14			7.45
15			8.19
16			8.62
17			9.29
18			9.69
19			10.26
20			10.89

Total	217
Median	4.02

larger frequency at item 4: "It provided the kind of experience I can apply to my own situation." Therefore, it might be better to report that a considerable number of persons responded that, "It provided the kind of experience that I can apply to my own situation," and that "It solved some problems for me."

Statements 5 and 7 were checked a number of times and these might be mentioned as well. Extreme scores, too, are interesting. There were six persons who said, "It was one of the most rewarding experiences I have ever had" (item 1). Six persons also said, "Exactly what I wanted" (item 2). But note that we do not know whether the same six checked statements 1 and 2. At the other extreme, only five checks reflected attitudes on the negative side of neutral, and this does not necessarily represent five different persons.

Under some circumstances, you may wish to ask respondents to arrange rankings, a series of options in rank order according to personal preference. When the number of things to be ranked is small and homogeneous, the ranking may force persons to make discriminations they would not otherwise make. For example, this section of the Guide contains a number of key concepts on selecting and developing evaluation instruments. An appropriate posttest might seek to find out which topics are the most helpful to evaluators. Consider two methods for trying to collect these judgments. The first is a ranking procedure; the second is a rating procedure.

. Which among the following topics did you find most helpful personally?

Method 1: Ranking Procedure

Directions: Rank order by assigning a "1" to that topic which was most helpful, a "2" to that section you found next most useful, etc.

_____1. reliability and validity
_____2. achievement tests
_____3. questionnaires
_____4. observational techniques
_____5. other behavior
_____6. sources of information about instruments
_____7. developing assessment instruments

Method 2: Rating Procedure

Directions: Rate on a 5-point scale each of the following major topics discussed in this section, using the following scale:

1 = Of no use: I will never need to know or use this.
2 = Of minimal use: I may have to use this information some time.
3 = Of some potential use: If I have to make use of this information, this topic will be helpful.
4 = Of considerable use: I expect I will need to use this information.
5 = Of maximum use: I will surely have to make use of this information.

	1 no use	2 min. use	3 some use	4 much use	5 max. use
1. Reliability and Validity					
2. Achievement Tests					
3. Questionnaires					
4. Observational Techniques					
5. Other Behavior					
6. Sources of Information about Instruments					
7. Development Assessment Instruments					

Different kinds of information are asked for in the two methods. The first method asks how each topic stands in relation to the other topics (norm-referenced approach). The second method asks how valuable each topic is in terms of its usefulness (criterion-referenced approach).

Actually, the two approaches could be combined, and both kinds of information obtained. The point is, the program evaluator must anticipate what kind of results are wanted by knowing beforehand how those results will be used.

Summary of Questionnaire Item Structure

Item Type	Advantages	Disadvantages
Open-Ended	- Free responses - Reasons can be given	- Difficult and time consuming to respond to, to score, and to interpret
Checklist	- Simple options - Easy to interpret	- Limited response - Only present/absent or yes/no responses
Multiple Choice	- Provides closure on questions - Simple options	- Limited response - Only correct or incorrect responses - Limited information on reason for judgment expressd
Ratings	- Degree of judgment identified; values assigned	- Directionality (-/+) confusing - No information on reason for judgment
Rankings	- Provide a norm-referenced approach - Easy to develop and use	- Can only be used with limited number of homogeneous topics

LEARNING EXERCISE 10: JUDGING ITEMS

"At the end of the eighth month of the school year, 70 percent of the participating students and parents will judge their reading program to have been successfully implemented, as measured by a questionnaire."

The evaluator accepts as valid this measure of students' and parents' attitudes about how much has been learned and whether enjoyment of reading and independent reading practice have increased.

The questions on pages 44 - 46 represent a preliminary list of those that may be used on the questionnaire. Your task is to critize each question and decide which ones should be included, revised, or discarded and to give the reasons.

In criticizing each item, consider the following three important criteria:

1. Appropriateness or Validity of the Item
   Is the item assessing something (knowledge, behaviors, attitudes, etc.) which indicates whether or not the evaluation objective has been reached?

2. Clarity of the Item
   Is the item written in a way that everyone will interpret it in more or less the same way? Is the item misleading or ambiguous? Does it present just one concept at a time?

3. Accuracy of the Response
   Do you think that the person answering the item will give an honest response or an accurate response? Is it emotionally loaded? Does it tend to bias or lead the respondent?

Study each question in relation to the above criteria. Decide whether you would accept it as is or revise or discard it. If you decide that any of the items should be revised or discarded, enter an "R" or a "D"in the appropriate columns on pages 44 - 46 and then state your reasons. When you have completed all 12 questions, go back and try to revise those items you decided should be revised, using the space on page D-46.

Student and Parent Questionnaire on Reading Program

Critique of Items

(If you decide to accept an item, check the "Accept" column. If you decide to revise or discard, enter "R" or "D" and write reasons in the next column. After finishing all question, enter revisions.)

Item	Accept	Revise or Discard	Reasons	Revisions
1. Did you learn anything from this year's reading program?				
2. Do you think the instructional methods of this class are better than in other classes you have taken?				
3. Did you like this year's program?				
4. As compared with the reading program taken last year, do you feel that this program was better?				
5. Do you think getting individual help is a good way to learn?				

Student and Parent Questionnaire on Reading Program

Critique of Items (cont'd)

(If you decide to accept an item, check the "Accept" column. If you decide to revise or discard, enter "R" or "D" and write reasons in the next column. After finishing all question, enter revisions.)

Item	Accept	Revise or Discard	Reasons	Revisions
6. Were the teaching assistants pleasant and helpful?				
7. Are parents pleased with the progress you have made in this year's reading program?				
8. Would you recommend this program to one who needs to improve his or her reading?				
9. If a more advanced class like this one was formed, would you want to be in it?				
10. Do you enjoy reading more now because of the program?				

Student and Parent Questionnaire on Reading Program

Critique of Items (cont'd)

(If you decide to accept an item, check the "Accept" column. If you decide to revise or discard, enter "R" or "D" and write reasons in the next column. After finishing all questions, enter revisions.)

Item	Accept	Revise or Discard	Reasons	Revisions
11. Do you read more books on your own than you did last year because of the program?				
12. In terms of how well you *now* read, do you think this year's program has something to do with that?				

ANSWERS
JUDGING ITEMS

(1) Item	(2) Accept	(3) Revise or Discard	Possible Reasons
1.		X	Too general: Learn "anything" could include "Yes, I learned to hate books."
2.		X	"Instructional Methods" is jargon which may have little meaning for 7th graders.
3.	X		
4.		X	Too general, vague and ambiguous. May encourage yes response for extraneous reasons.
5	X		
6.		X	Asks for two judgments in a single item ("Pleasant and helpful"). Pleasantness is not necessarily related to the learning process.
7.		X	Should be individualized to fit one student and his or her parents.
8.	X		
9.		X	"More advanced" may be ambiguous; highly competitive students may be more inclined to say yes than others; some students may reason that they did so well in the current class they really don't need more.
10.	X		
11.	X		
12.		X	May mean either the student thought his good results or his poor results could be attributed to the program.

Observational Techniques and Instruments

While both achievement tests and questionnaires can give valuable information for program evaluation, there are many kinds of information that cannot be obtained from them. Observational techniques and instruments for recording observations provide an added dimension. Like any other assessment instrument, there are both advantages and disadvantages in using them.

Advantages and Disadvantages of Observation

Advantages	Disadvantages
1. Can provide valid and reliable information on social-emotional-personal adjustment not possible with other traditional methods.	1. Difficult to get valid and reliable data.
2. Can test a person's ability to apply information in life-like situations.	2. Long period of training and experience may be required for the observer.
3. Easily adapted to a variety of tasks, settings, and individuals at all educational levels.	3. Many activities take place simultaneously in a classroom and it can be difficult to record behaviors that are significant.
4. Provides a valuable supplement to achievement data. 5. Can provide both qualitative and quantitative data.	4. Interpretation of observational findings must take into account the context, must not generalize from a too limited sampling of behaviors, must not give disproportionate weight to negative incidents, and must be as objective as possible, given the data at hand.

In program evaluation, observational techniques are most helpful in obtaining data on:

- Group participation and responsibility
- Individual student interaction with the group
- Teacher interaction with class

A wide variety of instruments may be used to record observations. Rating scales and checklists are commonly used. But anecdotal records, sociometric techniques, and highly developed systems, such as the Flanders System of Interaction Analysis,* are also ways of collecting observational data. Two examples follow:

Example 1: Observation and Analysis of Question-Answer-Feedback Sequences in Classroom Instruction

Suppose one objective of a program is to improve teaching techniques that encourage student participation in general classroom discussions. Suppose further that teachers have been told that the use of praise and affirmation of students, correct responses is to be preferred over negative or critical remarks about students' incorrect responses, and that directing a question to another student or rephrasing to make it easier are to be preferred over simply giving the class the answer.

The observation system shown on the following page could be used both before teachers receive instruction and afterward to determine the effectiveness of teacher training. The instrument for recording observation could be printed on both sides of 4" x 6" cards, as illustrated in Figure 1. Effective use of this instrument would require setting up a schedule for observing each teacher, both before and after they have received instruction. Several observations at periodic intervals after instruction might be scheduled. Important considerations include:

1. The amount of observation time should be the same for each classroom.

---

* Flanders, N.A. Teacher influence, pupil attitudes, and achievement. Cooperative Research Monograph No. 12, OE 25040. Washington, D.C.: U.S. Government Printing Office, 1965.

Figure 1. System for Recording Observations of Teachers' Reactions*

Coding Categories for Question-Answer-Feedback Sequences

STUDENT SEX

SYMBOL	LABEL	DEFINITION
M	Male	The student answering the question is male.
F	Female	The student answering the question is female.
STUDENT RESPONSE		
+	Right	The teacher accepts the student's response as correct or satisfactory.
±	Part right	The teacher considers the student's response to be only partially correct or to be correct but incomplete.
—	Wrong	The teacher considers the student's response to be incorrect.
0	No answer	The student makes no response or says he doesn't know (code student's answer here if teacher gives a feedback reaction before he is able to respond.)
TEACHER FEEDBACK REACTION		
+ +	Praise	Teacher praises student either in words ("fine," "good," "wonderful," "good thinking") or by expressing verbal affirmation in a notably warm, joyous, or excited manner.
+	Affirm	Teacher simply affirms that the student's response is correct (nods, repeats answer, says "Yes," "OK," etc.).
0	No reaction	Teacher makes no response whatever to student's response—he simply goes on to something else.
—	Negate	Teacher simply indicates that the student's response is incorrect (shakes head, says "No," "That's not right," "Hm-mm," etc.).
—	Criticize	Teacher criticizes student, either in words ("You should know better than that," "That doesn't make any sense—you better play close attention," etc.) or by expressing verbal negation in a frustrated, angry, or disgusted manner.
Gives Ans	Teacher gives answer	Teacher provides the correct answer for the student.
Ask Other	Teacher asks another student	Teacher redirects the question, asking a different student to try to answer it.
Other Calls	Another student calls out answer	Another student calls out the correct answer, and the teacher acknowledges that it is correct.
Repeat	Repeats question	Teacher repeats the original question, either in its entirety or with a prompt ("Well?" "Do you know?" "What's the answer?").
Clue	Rephrase or clue	Teacher makes original question easier for student to answer by rephrasing it or by giving a clue.
New Ques	New question	Teacher asks a new question (i.e., a question that calls for a different answer than the original question called for).

STUDENT SEX				STUDENT RESPONSE					TEACHER FEEDBACK REACTION										
NO.	M	F		+	±	—	0		++	+	0	—	— —	GIVES ANS.	ASK OTHER	OTHER CALLS	RE-PEAT	CLUE	NEW QUES.
1		✓		✓						✓									
2	✓			✓						✓									
3	✓						✓											✓	
4	✓			✓					✓										
5	✓			✓						✓									
6		✓				✓							✓						✓
7	✓			✓					✓										
8	✓			✓							✓								
9	✓			✓							✓								
10	✓						✓						✓	✓					
11																			
12																			
13																			
14																			
15																			

---

*Good, T. L., and Brophy, J. E. Looking in classrooms. San Francisco: Harper and Row, 1973, pp. 62 and 63. Reprinted with permission c 1975, Harper and Row, Publishers, Inc.

2. Set observation times that will be best for all classrooms. Avoid periods immediately preceding or following vacations or special events. Early Monday morning and late Friday afternoon should be avoided.

3. Where possible, assign classrooms randomly to different time blocks. Classes and teachers vary as the school day goes on. In classroom observation, you want to get a fair sampling of classroom climate across all classrooms.

4. If a number of different observers are used, be sure they are adequately trained in the observaton procedure and that inter-rater reliability has been checked (this is discussed in a later section).

Example 2: Interaction among Groups

The next example is an observational technique that was used in a national survey* of 13- and 17-year-olds as a measure of an objective dealing with the ability to apply democratic procedures on a practical level when working in a group. It demonstrates one way of measuring interaction among students and illustrates the need for very explicit directons in the training of observers and the recording of data.

Setting: (A group of eight students was asked to choose from a list the five most important issues between teenagers and adults, to rank order them according to importance, and to write a recommendation for at least the two most important problems, and for all five if they had time. They had 30 minutes to complete the task. The only rule was that a majority of the group must agree on anything they wrote. Two observers recorded individual acts of the group members as they discussed the issues, each observer recording different types of behavior. At no time did the observers participate in the discussion.)

---

* National Assessment of Educational Progress. Citizenship: National results. Denver, Colorado: National Assessment of Educational Progress November 1970.

List of Issues

Age 13

- Time Limits (for being home, in bed, etc.)
- Home Duties
- School Assignments
- Adult Books and Movies
- Sports and Other Activities
- Dating and Partiesctivities
- Parents' Approval of Friends
- Money (where from and how spent)
- Dress and Appearance
- Smoking
- Swearing
- Being Talked to Like an Adult

Age 17

- Censorship
- Curfew
- Voting Age
- Drinking
- Smoking
- Working Rules and Laws
- Marriage Rules and Laws
- Auto Insurance
- Dress and Appearance
- Military Service
- School Attendance
- Civil Liability
- Criminal Liability

The purpose was not to find out how students ranked issues but to observe the process by which they arrived at ranking decisions. Specifically, the behaviors to be looked for were:

- Took a clear position
- Gave a reason for a point of view
- Sought information related to the game from other team members or from the administrator
- Steered the task by organizing the group or by suggesting a change in procedure
- Defended the right of another group member to be heard or to hold a different opinion
- Defended own viewpoint contrary to a previous consensus
- Nontask behaviors

The recording forms were something like this:

School ______________________ Date ______________

Location ______________________ Time ______________

Observer 1

Gave a Reason	Took a Position	Opposed Group Alone (O) Yielded (Y) Convinced (C)
S	S	S
1.*	1.	1.
2.	2.	2.
3.	3.	3.
4.	4.	4.
5.	5.	5.
6.	6.	6.
7.	7.	7.
8.	8.	8.

*Each number identifies a given student.

School ______________________ Date ______________

Location ______________________ Time ______________

Observer 2

Steered Task	Sought Information	Defended Another	Nontask Action
S	S	S	S
1.*	1.	1.	1.
2.	2.	2.	2.
3.	3.	3.	3.
4.	4.	4.	4.
5.	5.	5.	5.
6.	6.	6.	6.
7.	7.	7.	7.
8.	8.	8.	8.

*Each number identifies a given student.

However, the general instructions to observers and the specific behaviors to be observed were more explicit:

General Instructions to Both Observers

1. Only overt actions are to be recorded, not general impressions
2. A single event or action may be scored in more than one category. Many comments made by group members will not be scoreable in any category, however.
3. When the group starts its task, observers should take positions in the background rather than as members of the working group, so that the group will not depend on the observers as moderators, leaders, etc. The observers must be seated close enough and in such a way that they can easily identify who is talking. So as not to be confused by numbering, both observers should probably sit where student No. 1 is at their immediate left.
4. Reliable observation can be maintained only by intensive effort and practice in use of the categories. Before each session, an observer should review carefully the categories he is to observe so that he can keep incisive definitions clearly in mind at all times. Tryouts have indicated that it is all too easy for the observer to err in two directions in particular: (a) The concept of the behavior category is loosened so that too many inappropriate behaviors are included; (b) in concentrating on certain categories, other categories are not attended to, and behaviors fitting those categories are thereby not included.
5. Whenever an indicated behavior occurs, the observer should make a check (✓) in the appropriate column on the line for the student who demonstrated that behavior. With the exception of the Oppose Group Alone category, a student is scored only the first time he demonstrates the behavior. Observers will find that some categories will be scored for most students quickly in the session. Observers should then focus their attention mainly on those categories not yet scored

(Continued on next page)

and on those students who speak infrequently, so as not to miss the rare times these categories and students will be scored. There is no need to give further attention to categories and students already checked (except instances of "Oppose Group Alone").

6. If in doubt whether or not a particular behavior should be scored, the observer should not score it. After each session, any confusions about scoring should be discussed between team members and the project director called if necessary to resolve a frequently occurring problem.

Each of the seven behaviors is given in as much detail as the general instructions. Examples of directions to observers for one of the seven behaviors to be observed follow:

Directions to Observer for Behavior

"Steered Task"

Score subjects in the "Steered Task" column on page D-45 for the kinds of behavior listed below. (Do not score nonverbal behavior which might seem to fit the category or an utterance you are in doubt about):

1. Attempts to organize the task for the group or attempts to change some procedure for accomplishing the task. (Do not score when S tries to steer the group toward an incorrect or irrelevant performance of the task, e.g.--"Let's not worry about writing anything down; let's just have a good discussion on these issues.")
2. Notes the need for organization or change in procedure. (Asking whether the proper procedure is or is not scored)
3. Notes the need for a chairperson.
4. Calls for a vote or notes the need for consensus.
5. Reminds others what the main task is or what the rules are. (Merely reminding others of their next step is not scored.)
6. Tries to stop others from cutting up or arouses drifters. (Merely attempting to quiet the group is not scored.)

(Continued on next page)

7. Notes that present discussion is on a tangent.
8. Notes time priorities and stresses the importance of time in completing the task.
9. Volunteers or agrees to write down task products or expresses the need for such a recording.
10. Tries to move the group on to the next step and gives a specific procedural reason for doing so. (Trying to change the topic or proposing a new topic to be discussed is not scored unless S gives a procedural reason for doing so such as lack of time. Don't score "What's next?")

For those who may be interested in results, this is what happened in the 1970 survey:

Results of 1970 Survey

	% Who Did This at Least Once	
	Age	
	13	17
Took a clear position	62%	67%
Gave a reason for a point of view	67	79
Sought information related to the game from other team members or from the administrator	54	55
Steered the task by organizing the group or by suggesting a change in procedure	51	39
Defended the right of another group member to be heard or to hold a different opinion	4	1
Defended own viewpoint contrary to a previous consensus	6	24

Cautions in the Use of Observation Instruments. Any time more than one person is involved in collecting data with an observation instrument, the program evaluator must be concerned with consistency of those data. Standardized instruments usually have specific directions and information on inter-rater reliability. Instruments that are not standardized probably do not have this feature, and the program evaluator must make his own provisions. In general, whether or not the instrument is standardized, it is good practice to use the following procedure:

Assuring Inter-Rater Reliability on Observation Instruments

1. Train raters in use of instruments.
2. Have raters use the instruments on a group similar to that they will be observing (field test).
3. Compare results of raters at field-test stage.
4. If results are not the same, discuss dissimilarities, and retrain raters, or revise instrument.
5. Repeat steps 2 through 4 until satisfactory results are obtained.

Selecting an Observation Instrument. When selecting an observation instrument, consider the following four steps:

1. Define the factors on traits that match the program/evaluation objectives. For example, if the objective stated that individualized instruction should occur in the classroom and that teachers should use a range of equipment and materials and aides, all these factors should be covered somewhere in the observation instrument.
2. Identify existing observation instruments and determine that they deal with those factors. For example, use the Simon and Boyer (1967) edited text Mirrors for Behavior, volumes 1-4 as a resource for identifying appropriate observation items and formats or make adaptations from existing instruments. (Anita Simon and E. Gil Boyer, Philadelphia, Research for Better Schools, 1967.) Or, refer to Good and Brophy cited earlier in this section.
3. Gauge the advantages and disadvantages of the instruments. Use questions such as the following to help assess the worth of an instrument:

Content Validity:

- What kinds of data can the evaluator collect when using the instrument?
- To what extent are the data going to provide the evaluator with the needed information?

Reliability:

- To what extent can the evaluator trust the data produced by the instrument?

Technical Information:

- Is there back-up statistical information?

Scoring:

- What kind of scores are generated?

Usability:

- How long does it take to administer?
- How much support equipment is required?
- Are the instructions easy to understand?
- How difficult is it to train someone in its use?

Cost:

- How much will it require in resources (time/money/personnel)?

4. Remember need for:

- Comprehensive set of instructions; and
- Training for the observer(s).

LEARNING EXERCISE 11: CRITICIZING A CLASSROOM OBSERVATION INSTRUMENT

Directions: Study the observation instrument on the next page and make judgments about its adequacy. On the sheet following the instrument, record your responses. Think about how you would use it and the kinds of information you would get from it. "Acceptable" means you think you could use it and get useful information. Consider the following:

1. Identifying information: In six months will you know where it came from?
2. Scale points: Are they well-defined and functional?
3. Directions for use: Is it clear how the observer proceeds?
4. Coverage: Are most important classroom variables included?
5. Clarity and scorability: Are items to be observed clearly specified and free of ambiguity?

Classroom Observation Instrument

Teacher____________________
Observer____________________

	RATING SCALE				
	Good	Adequate	Below Ave.	Poor	N/A
Students					
1. Students begin work with minimal teacher direction.					
2. Students concentrate on their own work with minimal distractions.					
3. Students seek out staff and other students for assistance.					
Staff					
4. Staff prepares materials in advance and is available before and after class.					
5. Staff interacts appropriately with students at their level, in conversational manner, and with enthusiasm.					
6. Staff operates in team-like manner and assists each other as needed.					
Room					
7. Classroom zones and areas are well-defined for students and staff.					
8. Classroom is comfortable (temperature, visual displays, physical arrangements).					
9. Physical space is efficiently used by staff and students.					
Materials					
10. Materials are clearly marked and available to students.					
11. Books and other materials are displayed to catch student interest.					
12. Adequate amount of materials is available for carrying out the program.					
Program					
13. Realistic student goals are encouraged and appear to be known by the students.					
14. Record-keeping procedures (attendance and student progress) are maintained and easily provide information to the staff all the time.					
15. Student programs are checked and modified as needed.					
16. Some evidence of the purpose and offerings of the program can be seen in the room or in the students' materials.					

Criticizing a Classroom Observation Instrument

RESPONSE SHEET

	Acceptable		If not acceptable, what is the reason?
	Yes	No	
1. Identifying information			
2. Scale points			
3. Directions for use			
4. Coverage			
5. Clarity			
6. Scorability			

Criticizing a Classroom Observation Instrument

ANSWERS

	Acceptable		If not acceptable, what is the reason?
	Yes	No	
1. Identifying information		X	Not enough space to write I.D.; no date given; layout unattractive
2. Scale points		X	No attempt to provide observer with frame of reference; what is "good", what is "poor"? Labels don't seem to fit with items to be observed.
3. Directions for use		X	None exist! How long does one observe and under what conditions? What is the purpose of this instrument?
4. Coverage	X		
5. Clarity		X	Too much is left to the interpretation of the observer. What is "minimal teacher direction?" How can you tell if space is "efficiently used?"
6. Scorability	?		Sensible scoring system could easily be devised, and directions included for applying it. Summary information could then be taken off the completed instruments for analysis.

Other Behaviors

Often an evaluator can gain access to data which are readily available and which do not require a formal data-collection instrument. Such data are called unobtrusive measures.

Use of these types of measures is appropriate if an evaluation objective suggests that specific changes are expected and if available information shows that these changes are occurring. For example, the attendance rate of students will increase; the use of the reading lab will increase; the grade-point average will improve; the number of cuts from a class will decrease; the number of discipline referrals to the principal's office will decrease, and so on. Data of these kinds can be useful in the evaluation design as additional indexes of program success.

Unobtrusive measures can also be used as indirect measures of attitudes and interests. For example, instead of asking a student, "Do you ever, of your own accord, read humorous stories or books of satire?", you might check the school library to see what the circulation records show for this category. To find out what science topics are most popular, you could look for pages in the encyclopedia of science that are worn, have thumbprints, or are dog-eared. To find out if a new unit or program is interesting to its participant, you could check absence records before it starts and periodically while it is in progress.

The greatest advantage of unobtrusive measures is that the data-collection procedures do not themselves influence the results. Students may behave differently when an observer is present or when they are taking a test or answering a questionnaire. The experience of taking tests itself may influence subsequent performance. But with unobtrusive measures, students are unaware that program-related data are being collected. As a consequence, their behavior in the program is unaffected.

Metfessel and Michael (1967) have compiled an extensive list of unobtrusive indicators of student behaviors. An abbreviated list* follows:

Indicators of Status or Change in Student Behavior Other Than Those Measured by Tests, Inventories, and Observation Scales in Relation to the Task of Evaluating Objectives of School Programs.

1. Anecdotal records and case histories: critical incidents noted including frequencies of behaviors judged to be highly undesirable or highly deserving of commendation
2. Attendance: frequency and duration when attendance is required or considered optional (as in club meetng, special events, or off-campus activities)
3. Autobiographical data: behaviors reported that could be classified and subsequently assigned judgmental values concerning their appropriateness relative to specific objectives concerned with human development
4. Citations: commendatory in both formal and informal media or communication such as in the newspaper, television, school assembly, classroom, bulletin board, or elsewhere
5. Extracurricular activities: frequency or duration of participation in observable behaviors amenable to classification such as taking part in athletic events, charity drives, cultural activities, and numerous service-related avocational endeavors
6. Grade placement: the success or lack of success in being promoted or retained; number of times accelerated or skipped
7. Performance: awards, extra-credit assignments and associated points earned, number of books or other learning materials taken out of the library, products exhibited at competitive events
8. Recidivism by students: incidents (presence or absence or frequency of occurrence) of a given student's returning to a probationary status, to a detention facility, or to observable behavior patterns judged to be socially undesirable (intoxicated state, dope addiction, hostile acts, sexual deviation)

---

* Adapted from Metfessel, N.W., and Michael, W.B. A paradigm involving multiple criterion measures for the evaluation of effectiveness of school programs. Educational and Psychological Measurement, 1967, 27, 931-943.

Other possible indicators include: absences, appointments kept or broken, assignments completed, changes in program or in teacher as requested by student, choices expressed or carried out, disciplinary actions taken, number of dropouts, elected positions held, grade-point average, grouping, homework assignments, leisure activities, library card possessed, numbers of units or courses carried, peer group participation, recommendations or other referrals, skills, social mobility, tardiness, transiency, and transfers and withdrawals from school.

Indicators of Status or Change in Cognitive and Affective Behaviors of Teachers and Other School Personnel in Relation to the Evaluation of School Programs

1. Attendance: frequency of, at professional meetings or at inservice training programs, institutes, summer schools, colleges and universities (for advanced training) from which inferences can be drawn regarding the professional person's desire to improve his competence
2. Mail: frequency of positive and negative statements in written correspondence about teachers, counselors, administrators, and other personnel
3. Memberships, including elective positions held in professional and community organizations; frequency and duration of association
4. Rating scales and checklists (e.g., graphic rating scales of the semantic differential) of teachers' behaviors in the classroom or of administrators' behavior in the school setting regarding changes of behavior in professional competence, skills, attitudes, adjustment, interests, and work efficiency
5. Records and reporting procedures practiced by administrators, counselors, and teachers; judgments of adequacy by outside consultants

Other possible indicators include: article written; grade-point average; load carried by teacher; moonlighting; nominations by peers, students, administrators, or parents for outstanding service and/or professional competencies; termination; request for transfers.

Indicators of Community Behaviors in Relation to the Evaluation of School Programs

1. Alumni participation: numbers of visitations; extent of involvement in PTA activities; amount of support of a tangible (financial) or a service nature to a continuing school program or activity; attendance at special school events, at meeting of the board of education, or at other group activities by parents
2. Conferences between parent-teacher, parent-counselor, parent-administrator sought by parents; frequency of
3. Letters (mail): frequency of requests for information, materials and service; frequency of praiseworthy or critical comments about school programs and services and about personnel participating in them
4. Participant analysis of alumni: determination of locale of graduates, occupation, affiliation with particular institutions or outside agencies
5. Parental response to letters and report cards upon written or oral request by school personnel: frequency of compliance by parents
6. Telephone calls from parents, alumni, and from personnel in communications media (e.g., newspaper reporters): frequency, duration, and quantifiable judgments about statements monitored from telephone conversations
7. Interview data

Even though no formal instrument is required, some device (logs or summary sheets must be devised to collect unobtrusive measures.

5. SOURCES OF INFORMATION ABOUT INSTRUMENTS

Buros' Mental Measurements Yearbook is the best-known resource for locating published assessment instruments, but certainly not the only one. The references listed here give fairly comprehensive coverage over a wide range of instrument types, except for criterion-referenced tests, which have been treated in some detail in the discussion on achievement tests. More complete references can be found in item VII, An Annotated Bibliography of Guides for Test Selection, of Section J in this Guide.

Where to Locate Information About Assessment Instruments

Source	Type of Information
Buros' Seventh Mental Measurement Yearbook	Critical review on currently published standardized tests
Buros' Tests in Print	Comprehensive test bibliography and index to first six Mental Measurements Yearbooks
Center for Study of Evaluation at University of California at Los Angeles (Hoepfner)	Ratings on validity, reliability, appropriateness, ease of administration, etc., on published standardized tests
Test Publishers' Catalogs	Newest materials (sometimes not found in Buros)
Tests and Measurements in Child Development (Johnson and Bommarito)	Experimental instruments in child development (self-concepts, attitudes, social behavior)
Socioemotional Measures for Preschool and Kindergarten children (Walker)	Descriptions of 143 tests and measures of social and emotional development (includes some technical information)
Measures of Social Psychological Attitudes (Robinson and Shaver)	Critical reviews of tests (mostly experimental) in 8 general categories: life satisfaction, self-esteem, alienation, authoritarianism, sociopolitical attitudes, values, general attitudes toward people, and religious attitudes
Mirrors for Behavior (Simon and Boyer)	Existing observation instruments
ERIC Clearinghouse on Tests, Measurement, and Evaluation (Educational Testing Service)	Has annotated bibliographies of tests in many areas: measures of social skills, measures related to school-based attitudes, self concept, educationally disadvantaged, assessment of teachers, criterion-referenced tests
ETS Test Collection	A library of some 10,000 tests and other measurement devices representing the instruments of all publishers. Access is based on guidelines of the American Psychological Association. Address specific inquiries by mail or telephone. A quarterly Test Collection Bulletin is available on a subscription basis
Professional Journals Educational and Psychological Measurement Journal of Educational Measurement Journal of Counseling Psychology Personnel and Guidance Journal	Reviews and validity studies of recently published or revised tests

Use Multiple Measures Whenever Possible

It is sometimes easy to shoot down a single measure of student achievement by discrediting its score for some technical reason. Credibility of program evaluation may be enhanced by the use of several measures of program effectiveness. The inclusion of attitudes of parents, students, and staff as well as unobtrusive measures of student behavior will broaden the base of information from which judgments can be made.

Summary of Factors to Consider in Evaluating and Selecting Assessement Instruments

1. Reliability	Does the instrument give the same results when repeated?
2. Validity	Does the instrument measure what it says it measures? Does the content match your program objectives? Is it free of bias for different subgroups?
3. Content	Are the items related to program objectives
4. Administration Mode and Time	Is the instrument adminstered in groups or individually, by interview or observation? What qualifications does the administrator need? Are directions for administration adequate? Is equipment required for administration available? Is time required reasonable for the results expected?
5. Scoring	Is scoring by hand or by machine? Are directions for scoring adequate?
6. Format and Interest	What is the general editorial quality? Will it hold the test taker's interest, and are directions easy for test takers to understand?
7. Scores and Norms	If normed, what are the characteristics of norm groups? When was it normed? Are interpretive aids available?

LEARNING EXERCISE 12: SELECTING NORM-REFERENCED TESTS

Directions: Select and check one of the four objectives listed below, then consult the descriptive information on D-71 - D-72. Select up to three instruments* you think would be appropriate for that objective and list them on the following page. Record a "yes" in those boxes in which answers to the questions seem to be affirmative.

Objectives

1. Second- and third-grade students participating in the bilingual-bicultural program will have a mean score of 20 or higher on the ______________________________ series of tests of cultural similarities and differences. One test will be given after completion of each cultural unit.

2. The median percentile rank in reading comprehension for third-grade students participating in the remedial program will be eight points higher on the posttest given in May than on the pretest given to the same students in October. The test to be used is ______________________ test.

3. Kindergarten children at School Z with a 75 percent or better attendance will show nine months gain or more in language usage on the ______________ language test after nine months of instruction.

4. All tenth grade students receiving remedial math instruction will show at least a five-month mean gain in math computations for every five months of instruction. Gain will be measured by the ____________ test.

---

* At page D-73, judgments by specialists about the listed tests are shown.

Selecting Norm-Referenced Tests

Criteria	Instrument 1 Name________	Instrument 2 Name________	Instrument 3 Name________
1. Is the instrument a valid measure?			
2. Is the instrument a reliable measure?			
3. Is the instrument appropriate to use on the population to be assessed?			
4. Does the instrument yield objective data?			
5. Is the instrument easy to administer and score?			
6. Are minimum time and resources required to administer and score the instrument?			
7. Is the administration of the instrument nondisruptive to classroom learning activities?			
8. Will the instrument provide data which are useful for decision making at both the classroom level and the progam-administrative level?			
9. Is the cost of the instrument reasonable and within budgetary constraints?			

BRIEF DESCRIPTION OF TESTS

American School Achievement Tests by Robert V. Young, et. al.; Level - Primary I (Grade 1), Primary II (Grades 2-3), Intermediate (Grades 4-6), Advanced (Grades 7-9); Forms A,B,D,E; 1955-59 (BMC)

Subtest: Primary I: Word recognition, word meaning, numbers. Primary II: Sentence and word meaning, paragraph meaning, computation problems, language usage, spelling.

Intermediate: Sentence and word meaning, paragraph meaning, arithmetic computation, arithmetic problems, language, social studies, science.

Advanced: Sentence and word meaning, paragraph meaning, arithmetic computation, arithmetic problems, language, spelling, social studies, science.

CIRCUS, anonymous; Age: 4-5 years; 1972 (ETS)

Subtest: What words mean, how much and how many, look-alikes, copy what you see, finding letters and numbers, noises, how words sound, how words work, listen to the story, say and tell, do you know, see and remember, think it through, make a tree, activities inventory, teacher questionnaire, test-taking behavior

Comprehensive Tests of Basic Skills, anonymous; Level - I (Grades 2.5-4), Level II (Grades 4-6), Level III (Grades 6-8), Level IV (Grades 8-12); Forms Q, R & S; 1968, 1973 (CTB) Form S only: A & B (Grades K-1), C (Grades 1.5-2).

Subtest: Reading, language, arithmetic, study skills.

Cooperative Primary Tests, anonymous; Level: Grades 1-3; Form B; 1965 (ETS).

Subtest: Reading, listening, word analysis, mathematics, writing skills.

Durrell Listening Reading Series by Donald D. Durrell, et. al.; Level - Primary (Grades 1-3.5), Intermediate (Grades 3.5-6), Advanced (Grades 7-9); Form DE; 1969 (HBJ).

Subtest: Vocabulary listening, paragraph listening, vocabulary reading, paragraph reading

Gates-MacGinitie Reading Tests by Arthur Gates, Walter MacGinitie; Level - Primary B (Grade 2), Primary C (Grade 3), Survey D (Grades 4-6), Survey F (Grades 10-12); 1964-69 (BEM)

Subtests: Primary levels: Vocabulary and comprehension

Survey levels: Speed, accuracy, vocabulary and comprehension

Metropolitan Achievement Tests, 1970 Edition by Walter N. Durost, et. al.; Level - Primary I (Grades 1.5-2.4), Primary II (Grades 2.5-3.4), Elementary (Grades 3.5-4.9), Intermediate (Grades 5.0-6.9), Advanced (Grades 7.0-9.5); Form F & G; 1959 edition, Forms A and B also available (HBJ)

Tests: Primary: Listening for sounds, reading, numbers. Primary I: Test 1 - Word knowledge, Test 2 - Word analysis, Test 3 - Reading, Test 4 - Mathematics. Primary II: Test 1 - Word knowldege, Test 2 - Word analysis, Test 3 - Reading, Test 4 - Spelling, Test 5-7 - Mathematic computation, concepts, problem solving.

Elementary: Test 1 - Word knowledge, Test 2 - Reading, Test 3 - Language, Test 4 - Spelling, Test 5-7 - Mathematics computation, concepts, problem solving

Intermediate & Advanced: Test 1 - Word knowledge, Test 2 - Reading, Test 3 - Language, Test 4 - Spelling, Test 5-7 - Mathematics computation, concepts, problem solving, Test 8 - Science, Test 9 - Social studies.

Stanford Early School Achievement Tests by Richard Madden and Eric Gardner; Level - I (Grades K.1-1.1), II (Grade 1.0-1.8); 1969 (HBJ)

Subtests: The environment, social studies and science, mathematics, letters and sounds, aural comprehension, word reading, sentence reading

Tests of Basic Experiences by Margaret H. Moss; Level - K (Preschool-K); L (Grades K-1); 1970 (CTB)

Subtests: General concepts, mathematics, language, science, social studies (Also, Spanish directions, supplement to manual available)

Judgments of Tests

RATINGS* Good Questionable Poor	Is the instrument a valid measure?	Is the instrument a reliable measure?	Is the instrument appropriate to use on the population to be assessed?	Does the instrument yield objective data?	Is the instrument easy to administer and score?	Are time and resources required to administer and score the instrument minimum?	Is the administration of the instrument nondisruptive to the classroom learning activities?	Will the instrument provide data which is useful for decision making at both the classroom level and the program administration level?	Is the cost of the instrument reasonable and within budgetary constraints?
American School Achievement Tests Primary I, II	Questionable	Questionable	Questionable	Questionable	Good	Good	Good	Questionable	Good
Circus Level 4-6 Years	Good	Good	Good	Good	Questionable on scoring	Questionable	Questionable	Good	Good
Comprehensive Tests of Basic Skills, Forms A,B,C	Good	Good	Good	Good	Questionable if scored by hand	Good	Good	Good	Good
Durrell Listening Reading Series	Good	Good	Good	Good	Questionable on administration	Questionable	Good	Good	Good
Gates MacGinitie Reading Test	Good if matched with content taught	Good	Good	Good	Good	Good	Good	Good	Good
Metropolitan Achievement Test	Good	Good	Good	Good	Questionable	Good	Good	Good	Good
Readiness Skills Test	Good	Good	Good	Good	Questionable	Questionable	Questionable	Good	Good
Stanford Achievement Test	Good	Good	Good	Good	Questionable	Good	Good	Good	Good
Stanford Early School Achievment Test	Good	Good	Good	Good	Questionable on scoring	Questionable	Questionable	Good	Good
Test of Basic Experiences Level K-L	Good	Good	Good	Good	Questionable on scoring	Questionable	Questionable	Good	Good

---

* Ratings were given by only two person with experience in this field. The ratings may differ greatly when reviewed by a larger number of persons.

LEARNING EXERCISE 13: SELECTING APPROPRIATE INSTRUMENTS

Directions: Abbreviated portions of objectives are listed below. Decide what type of measuring instrument would be most appropriate to use for each. Record the letter representing that type.

Instrument Type

A. Norm-Referenced Test
B. Criterion-Referenced Test
C. Questionnaire
D. Observation Record
E. Log
F. Not enough information is provided to make a decision.

_____ 1. Parents of participating compensatory education students will have positive attitudes. . .

_____ 2. Twenty-five percent of the students will achieve one standard deviation above the national mean. . .

_____ 3. Students will interact in a positive social manner during class activities. . .

_____ 4. Given a list of South American countries, students will be able to list the capitol of each country.

_____ 5. The number of discipline referrals to the principal will be reduced by 50 percent.

_____ 6. Students will check out books in category "A" more frequently than books in category "B". . .

_____ 7. At the end of the semester, students in the values clarification class will exhibit positive atitudes toward their parents' ethnic background. . .

_____ 8. At the end of the inservice workshop, teachers will be able to answer correctly 8 out of 10 cognitive questions based on content of the workshop. . .

_____ 9. Teacher effectiveness in promoting student interaction will increase. . .

_____ 10. The majority of parents with students attending school "X" will be aware of the auxiliary services available through the school. . .

ANSWERS

1. C
2. A
3. D
4. B
5. E
6. E
7. C
8. B
9. D
10. C

6. LOCATING EXISTING ASSESSMENT INSTRUMENTS VS. DEVELOPING ASSESSMENT INSTRUMENTS LOCALLY

Examples of every type of instrument discussed in this section exist somewhere. The program evaluator who spends time searching for available instruments that will meet his needs usually will be far ahead of the one who decides to launch a school-wide or district-wide effort to develop tests, questionnaires, or observation records locally. The development of good assessment instruments is a much more exacting and demanding task than is often realized. Questionnaires and observation schedules can usually be adapted to local needs. But even adaptation takes care and thought. Criterion-referenced tests that can be assembled for your purposes from existing items are becoming increasingly available through commercial sources. The selection of any type of instrument must, of course, be done with care.

Developing Instruments

Unless you have highly trained technical staff and sufficient time and money, the development of instruments locally should only be undertaken as a last resort. Time and costs will vary with the magnitude of the job. Development of a reasonably straightforward achievement test should take eight months to a year to develop and a year to review, field test, and revise. The adaptation of an existing instrument for local use may be done in considerably less time, but even so, field test, review, and revisions steps should be given time to run their course.

People who write items, whether for achievement tests, questionnaires, observation, or other measures, must know the content area to be measured and must know basic techniques for item construction. In most cases, it is easier to train a content person in the art of test construction than to take a professional item-writer and teach him/her the content area. However, it is sometimes advisable to get persons with the two separate skills and have them work as a team. In any case, the content person should know the basics of good item construction. There are easy-to-follow rules in any basic test in measurement and evaluation. (See selected bibliography at the end of the Guide starting at I-1.)

The development of an evaluation instrument begins with a plan that specifies the information wanted. Determine what questions you want answers to. Program objectives serve as the basis for planning measurement needs.

Items are then written. It is good practice to develop more items than will be needed in the final instrument because some will be lost through review and field testing.

One of the greatest difficulties in constructing objective items is in getting plausible options. In the case of achievement tests, there must be one and only one best answer. The other options (or distractors) should be plausible answers for persons who do not know the right answer. In the development of questionnaire items, it is often impossible to anticipate all appropriate choices a person could make. A preferred procedure is to first administer items either in open-end or partial open-end format and use returns from field tests to set options. Alternatively, an item writer might try to guess what these options or choices are. For example, if you are designing an item to determine professional growth among staff members, you might guess at some possible activities, and then allow for an "Other" response.

Example: 1. In which of the following professional activities have you participated in the past year?

______ Enrolled in college or university course
______ Attended special workshop sessions
______ Observed in other classrooms
______ Done independent reading
______ Consulted with specialist
______ Other ______________________________

For achievement test items, a completely open-ended format should be used if there is any doubt about being able to anticipate good distractors. In such cases, the most frequently given wrong answers provide the best distractors.

Field testing should include all concerns related to the collection of data:

1. Are there adequate procedures for training persons who will collect the data?
2. Are the directions for administration clear and understandable?
3. Does the instrument itself give the kind of information you are seeking?
4. How long does test administration take?
5. Does the scoring key work?

The answers to such questions come from various sources and include both "hard" and "soft" data. There are statistical procedures for analyzing the data from the instrument itself, but getting oral responses by interviewing participants (both data collectors and persons tested) in a field test may also be necessary. This may be by group interview and should cover such points as clarity of tasks, ambiguities in individual items, and the actual mechanics of data collection.

Obviously, the group on whom you do the field test should be similar to the groups on whom you expect to use the instrument, but should not consist of members of that group.

Evaluators who find they must produce locally developed tests are strongly advised to seek help from measurement specialists. The essential steps in instrument development are outlined below.

Instrument Development Procedure

Activity	Questions to be Answered and Cautions
Develop plan.	What information do you want?
Prepare draft.	Developers should be able to write well and have knowledge about program content.
Have several persons review the first draft.	Are there ambiguities, omissions, and unnecessary pieces of information requested?
Develop directions for administration.	Try to assure that data will be collected under standard conditions.
Develop scoring key	Can each part be scored, and is there agreement on the scoring?
Field test and prepare draft including the directions for administration.	Administer it to persons like those on whom you plan to use it; ask them to criticize it; get time estimates.
Revise the instrument.	What went wrong? Fix it.
Repeat review, field-test, and revision steps as necessary.	Are you satisfied that the information obtained from this instrument will answer the originl questions you wanted to have answered in the first item above?

The preceding are management steps and do nothing to assure technical adequacy. Validity studies can be planned which use an outside criterion you believe to be independent of what you are trying to measure with your instrument. For example, if you want to validate a test of reading comprehension, you might do any or all of the following:

1. Ask teachers to rank students in order of their respective competencies in reading comprehension.
2. Ask parents how well their children can read and understand newspaper items.
3. Ask teachers to estimate what proportion of their students will pass each item on the test.

Checking for reliability is more involved and beyond the scope of the Guide.

7. REVIEW

As you have seen, there are several kinds of instruments that can be used for program evaluation. Which one(s) you select will depend to a great extent upon the nature of the evaluation design and the kinds of information that need to be collected.

Here is a review of what we have covered in this section of the Guide:

1. It is important to match instruments to program objectives.
2. Multiple measures for each program objective are desirable.
3. Different techniques and various types of instruments can be used.
4. There are both advantages and disadvantages to using different instruments and data-collection methods.
5. There are many sources of information on existing instruments.
6. It is generally better to adapt or adopt existing instruments than to develop new ones locally.
7. The development of adequate instruments locally is a costly, time-consuming, and demanding task.

SUMMARY OF BASIC EVALUATION INSTRUMENTS

	Types of Evaluation Measures	Categories of Instruments	Primary Structure of Items	Primary Kinds of Scores
PROGRAM OBJECTIVES	Achievement	Norm-Referenced Tests Criterion-Referenced Tests	. Objective . True/False . Multiple Choice . Matching	Raw Scores Grade Equivalents Percentiles Standard Scores Stanines Percentages
	Attitude	Questionnaires	Open-Ended Objective . Yes/No . Multiple Choice . Ratings Mixed	Categories Frequencies Percentages Ratings
	Interaction	Observation Record Forms	Open-Ended Objective . Ratings . Present/ Absent	Categories Frequencies Percentage Ratings Time
	Other Behaviors	Logs . Referral Reports . Attendance . Cuts . Grades . Diaries	Open-Ended Objective	Frequencies Categories

PROGRAM EVALUATOR'S GUIDE

Section E

COLLECT THE DATA

The Evaluation Improvement Program

PRECIS

After available instruments have been selected or plans made for developing new ones, the evaluator must then plan how to collect the data. Collecting data is a sensitive part of program evaluation because its success depends so much on the cooperation and often the hard work of others who are not connected with the evaluation team. Moreover, the logistics of moving evaluation instruments and data from place to place and the complexities of schedules involving hundreds of people present challenges not encountered in other parts of the program-evaluation process.

Planning for data collecting involves: specifying the subpopulations that are to serve as sources of information, deciding on who will be responsible for collecting the information, and deciding whether the collection will be carried out on an individual or a group basis. Special arrangements need to be planned to follow up when people are absent from group sessions or when individuals do not return questionnaires.

CONTENTS

Page

1. INTRODUCTION E-1

2. BASIC CONSIDERATIONS E-1
    Arrangements with School/Program Personnel E-2
    Personnel Who Collect the Data E-3
    Training Needed for Those Who Will Collect the Data E-4
    Time Schedule E-5
    Monitoring the Data-Collection Process E-5

3. THE MECHANICS OF DATA COLLECTION E-6
    Collection of Data from Groups of Persons E-6
    Collection of Survey Data E-10

LEARNING EXERCISE 14: PLANNING FOR DATA COLLECTION E-12

4. MONITORING AND RECORDING DATA ON PROGRAM ACTIVITIES AND CONDITIONS E-15
    Formative Process Evaluation E-15
    Formative Context Evaluation E-15
    Procedures for Collecting Process and Context Data E-16

1. INTRODUCTION

No matter how reliable and valid an instrument is, its usefulness can be completely destroyed by carelessness in the collection and handling of the data.

The problems that can arise during the data-collection stage are many and varied. Among some of the more common are these:

- Reading scores obtained the day before a vacation may not be comparable to those obtained a week earlier.
- Responses in an interview situation may be influenced by the race, sex, or status of the interviewer.
- Interrruptions or faulty directions can "destandardize" a standardized test.
- Voluntary responses to a mailed questionnaire may not be representative of the total population.

In addition, any data-collection plan must take into account a variety of logistical problems, the importance of which increases geometrically as the number of students, teachers, and schools involved in the program evaluation increases.

2. BASIC CONSIDERATIONS

Some of the most important considerations in planning for data collection concern:

- arrangements with school/program personnel
- personnel who will collect the data
- training needed
- time schedule
- monitoring the total data collection process

Arrangements with School/Program Personnel

In the several stages of collecting data, the evaluator needs the cooperation of school personnel. For example:

- The evaluator may need to obtain from the school certain records and lists of students, classrooms, teachers, holidays, faculty meetings, materials, etc., to select appropriate instruments and populations to be included in the program evaluation and to carry out other plans.
- The evaluator may need to train school personnel in the administration and use of evaluation instruments (e.g., standardized tests, questionnaires, surveys, checklists, etc.).
- The evaluator will need permission and active cooperation in collecting data in the school.

The kind of cooperation received often depends on how aware school personnel are of the benefits for students, teachers, and administrators expected to result from the program evaluation.

If the evaluator is able to explain the purpose of the evaluation and how feedback from it can be used, school personnel will respond more cooperatively. This understanding is the key to obtaining cooperation.

There are several ways to create a favorable attitude toward evaluation. The evaluator could convince the principal of the school to reserve a column in the school's newsletter for periodic reports on the program and its evaluation. The evaluator might meet with the participating teachers to discuss: the expected outcomes of the evaluation, the types of information needed and why that information is important and necessary, and what they will get for their efforts. The evaluator might also make personal contact with some parents of participating students to brief them and to help organize a parent-evaluation committee to help disseminate information related to evaluation activities.

Personnel Who Collect the Data

One of the decisions to be made in planning for data collection concerns those who will do the actual collecting. Should they be teachers? Secretaries? Students? Should they be people from inside the program or someone not involved in the program? How much working time will be required? How many people?*

In a program evaluation involving large-scale standardized testing, all teachers may be assigned to administer tests in their classrooms simply because they are the only available staff resources and physical facilities. In a smaller program evaluation, perhaps tests would be administered only by a small team of teachers or teams of teachers and aides. In any case, test administrators will need ample orientation and training.

The evaluator must also decide whether the persons collecting data should be from the program being evaluated or from outside the program. Some authorities have suggested that data collection should be carried out by outsiders unfamiliar with the objectives of the program to bring a totally unbiased viewpoint to the testing situation. The use of outsiders may serve to prevent two types of bias:

1. Program personnel may have vested interests in the program. They may tend to focus on those aspects of the program which are successful. They may interpret the data in a favorable manner, whether justified or not.

2. Program personnel may be particularly attentive to program objectives and might plan their evaluation accordingly, overlooking effects of the program that are observable but unanticipated.

   EXAMPLE: Programmed instruction in mathematics often has the side effect of improving reading ability. Narrow program evaluation would focus on mathematics achievement and might not uncover the positive effect on reading achievement.

---

*A statewide survey in California conducted in 1975 by Evaluation Improvement Project staff showed that most program evaluation data are collected by classroom personnel.

Training Needed for Those Who Will Collect the Data

The evaluator should assess the type and amount of training required to administer the different evaluation instruments and plan accordingly.

In the case of administering a standardized test, the basic requirements for data collectors are a willingness and ability to follow directions precisely. Training for the task will probably not need to be extensive--a brief orientation supplemented by a checklist for use during test administration. However, do not underestimate the need to periodically remind test administrators of what good testing conditions consist of (proper room environment, techniques of distributing and collecting materials, monitoring students, noting conditions that may invalidate a student's answers). And do not make the mistake of thinking that anyone can administer a standardized test. Be sure to specify what should be done if a person is absent.

If the evaluation involves the interviewing of parents using an interview guide, the decision as to who will administer that instrument is more critical than for a standardized test because the validity and reliability of the data collected may be substantially influenced by the personality of the interviewer. Training in this case would be more extensive.

When observation instruments are used, as in our reading program example, the evaluator needs to carefully study the instrument, arrange several pilot observations in a nonparticipating school, and possibly train a second person to participate in the pilot observations for purposes of comparing the data collected by two observers.

Plans should be made for both training and practice by observers. This is especially true with instruments developed locally. When observation instruments require judgments, some check of validity and reliability of the data collectors' ratings should be made during pilot situations. This should be done when the instrument is being developed and early enough to allow changes to be made prior to use.

Thus, the training of those who will participate in the collection of data is very important. In effect, if the data collectors are not carefully trained, the evaluator may have no data to analyze.

Time Schedule

The schedule for data collection will be partly determined by the evaluation design and by the deadlines for analyzing and reporting results. It will also be influenced by other factors ranging from school calendars and vacation days to curriculum plans and grading periods. The schedule should be as detailed as possible including such things as training sessions, testing-room preparations, space adjustments, materials delivery, as well as the actual data-collection activities. Do not schedule testing sessions just before or after holidays or in close proximity to major school events.

The evaluator should coordinate, prepare, and issue a data-collection schedule for each evaluation event. This schedule should indicate at least the following information:

1. WHEN the data are going to be collected (e.g., dates: October 22 from 10:00 a.m. to 11:45 a.m.)
2. WHERE the data are to be collected (e.g., Classroom A)
3. WHO is going to collect the data (e.g., the names of the persons responsible at each location)
4. WHO is to be evaluated, (e.g., names of the students or others)
5. HOW the data will be collected (e.g., name of the instrument[s] to be used)

Monitoring the Data-Collection Process

And finally, there must be assurances that the plans are carried out as expected, instruments are applied properly, data are recorded accurately, absentees are accounted for properly, and that steps are taken to correct or note untoward incidents which might bias the results. Monitoring can help assure that factors are measured in the ways you intended to measure them.

Procedures that are carried out carelessly during data collection may result in the measurement of extraneous factors such as the clarity of the directions given at the start of a test rather than of what was learned as a result of the program. Failures in data collection can jeopardize the entire program-evaluation effort.

Group Discussion

What experiences have you had in collecting data that could help others? What are the little things that can trip you up? Materials not arriving in time? Wrong materials distributed or not enough materials? Has security been a problem?

How would you have prevented the situation described in the ancedote below?

> The principal of a small elementary school noted that the answer sheets from one particular class were only about half filled. A large portion of the class had completed only about half the questions. Upon questioning the teacher who had administered the test, the principal found that a stop watch had been used to time each section of the test. However, this particular stop watch had a sweep-second hand that revolved twice for each minute. The teacher had inadvertently read one revolution (30 seconds) as one minute. Thus, the entire test had been administered in half the time it should have taken.

3. THE MECHANICS OF DATA COLLECTION

Collection of Data from Groups of Persons

Most data on students are gathered in a classroom setting (or in large units such as the school cafeteria or auditorium). When data are gathered under such concentrated conditions, it is far easier to control the situation and get valid data than it is in a survey. Organization and attention to detail are the keys.

Except in the smallest of schools, the collection of data from groups requires cooperation among people and coordination of activities by many persons. The respective roles of the program evaluator and the classroom teacher (or other persons responsible for administering tests and questionnaires or collecting observational data) must be clearly recognized and combined if valid data are to be gathered.

The following Checklist for the Program Evaluator and Sequences of Activities for the Data Collector specify the steps which must be taken in implementing an effort to collect data from groups.

Checklist for the Program Evaluator:
Collection of Data from Groups

______ 1. Instruments have been delivered to those administering them well in advance of the time they are to be used.

______ 2. Quantities, levels, and forms of instruments have been checked against actual needs.

______ 3. Storage in secure places has been arranged.

______ 4. Instruments and accompanying manuals and other materials have been thoroughly reviewed so that data collectors can be trained effectively.

______ 5. Persons to administer data-collection instruments have been carefully selected and trained and provided with their own sets of materials in advance.

______ 6. A detailed schedule has been prepared and distributed.

______ 7. Classroom distribution and collection procedures have been carefully worked out.

______ 8. Data collectors have been instructed what to do about absentees during testing periods.

______ 9. Specifications and arrangements for scoring have been made.

______ 1[illegible]. If scoring is to be done by an outside agency, answer sheets have been checked for completeness and organized for processing.

______ 11. Test books and all other materials have been returned to secure storage after use.

Sequence of Activities for the Data Collector

Before data collection:

1. Study the directions for administration, examine assessment instruments and answer sheets.
2. Rehearse process of administering instruments.
3. Clear up any potential problems with the program evaluator.

During data collection:

1. Prevent disruptions from outside sources (a TESTING--DO NOT DISTURB sign is recommended). Make sure room environment is comfortable.
2. Make announcements slowly and clearly. Try to motivate participants without causing anxiety.
3. Be sure each person has all materials and equipment needed.
4. Allow sufficient time for each person to fill in required identifying information (for young children, the data collector may need to do this step in advance).
5. Use exact wording given in printed instructions. Do not improvise or use short cuts unless directions for administration allow for variation. Be sure each person understands what he or she is to do.
6. Once testing begins, walk around the room to be sure everyone is working. Do not answer questions related to test content. ("Do the best you can" or "skip that one and go on to the next" can be used as a response, if necessary.)
7. Stop immediately when time is up.

After data collection:

1. Collect answer sheets first, then booklets.
2. Count all materials to be sure none is missing.
3. Alphabetize and check papers against group roster.
4. Check all papers for completeness of identifying information.
5. Prepare an exceptions list. (Did anyone become ill or leave the room during the session? Was there an unexpected fire drill during the session?) Any condition that could potentially invalidate the results of one person or the entire class should be noted.
6. Arrange and organize for scoring.

Collection of Survey Data

When data are collected at a specific time and at a specific place (i.e., in the classroom, after an inservice training session, at a meeting of the parents' groups), the program evaluator has greater control over conditions during data collection. However, not all data can be collected in this manner. The field survey, for example, has become one important method of obtaining information--particularly from parents and community groups. The conditions under which data are collected from a mailed survey are subject to very little control by the program evaluator. However, there are a few things the evaluator can do that may help increase the percentage of returns and usefulness of responses.

1. Keep the survey form short and to the point. Ten-page questionnaires often go into the round file.
2. Make your cover letter of instructions clear and concise. Explain why this information is important and why the recipient of the letter was included in the survey. Motivation is critical in getting responses to mailed surveys.
3. Encourage the respondent to complete the questionnaire at one sitting and when he or she is free of interruptions.
4. Follow up the questionnaires to increase returns. It takes 75 percent to 80 percent to give you unbiased results.

This last point is so important that it calls for a more detailed discussion.

How do you obtain a high percentage return on surveys? A high rate of return on surveys comes from a high level of effort. Mailing survey forms or sending them home with students and waiting for the responses to come in does not involve a high level of effort. Most persons who are going to respond will do so within the first two weeks after receiving the forms. Following up on them is essential.

The follow up may be done using a number of techniques, but generally these:

1. Notices in the press and on radio during the first two weeks
2. Sending a reminder and another copy of the questionnaire to nonrespondents near the end of the second week
3. Organizing some system of personal contact to try to reach nonrespondents and invite their cooperation

To take on the job of personally contacting all nonrespondents is generally not practical for the program evaluator. Consider what the other resources are. Are there school personnel who can be given lists of persons to call? Are there volunteer parent, student, or community groups whose aid can be enlisted? Are there "telephone trees" already organized by parent groups that might be used?

The amount of follow-up effort that can reasonably be expected is related to the number of persons in the original sample and, in turn, the number of nonrespondents who must be contacted. This argues for selection of the smallest sample feasible to achieve the desired result.

If the number of persons to be surveyed is not prohibitively large and if the evaluator can organize and train a small cadre of persons, a straight telephone survey rather than a mail survey could be planned. Each telephone interviewer would fill out a structured interview guide for each call made. This generally gets quicker results and a higher response rate. However, it may also mean making phone calls at night for a large portion of persons surveyed.

LEARNING EXERCISE 14: PLANNING FOR DATA COLLECTION

Directions: Assume you are the evaluator for a program and the decision has been made that evaluation data will be collected from students, staff, and parents, using a variety of assessment instruments. If possible, put this in the context of an ongoing program you have planned for next year. Columns (1) and (2) on the next page give general background on the types of instruments that have been selected and the populations from whom the data are to be collected.

In filling out the Plan for Data Collection on Page E-13, consider the questions:

Column (3): Who collects the data? -- classroom teachers, program evaluator, other designated staff members, volunteer parents, independent observer, other

Column (4): Will the data be collected from the target population in groups or will this be done individually?

Column (5): What will you plan to do about follow up? If persons are absent from a group session or do not respond to a survey, what provisions will you make?

Plan for Data Collection

(1) Type of Instrument	(2) Population on Whom Data Are To Be Collected	(3) Person Responsible for Collecting Data	(4) Group or Individual Activity	(5) Type of Follow up
1. Standardized Achievement Test	students			
2. Questionnaire	staff			
3. Questionnaire	parents			
4. Classroom Observation Scale	students			
5. Attendance Log (inservice training)	staff			

ANSWERS

(1) Type of Instrument	(2) Population on Whom Data Are To Be Collected	(3) Person Responsible for Collecting Data	(4) Group or Individual Activity	(5) Type of Follow up
1. Standardized Achievement Test	students	classroom teacher	group	test absentee at later date
2. Questionnaire	staff	program evaluator	group* individual	contact absentee contact non-respondent
3. Questionnaire	parents	program evaluator or designated staff member or volunteering parent teams	individual	contact non-respondent
4. Classroom Observation Scale	students	independent observer or classroom teacher	group	none
5. Attendance Log (inservice training)	staff	program evaluator or designated staff member	group	none

---

[*] Preferred, if convenient, as at a staff meeting.

4. MONITORING AND RECORDING DATA ON PROGRAM ACTIVITIES AND CONDITIONS

One of the important tasks of the evaluation process is the monitoring of program activities and conditions that affect the implementation of those activities. If the program staff does nct carry out the activities as planned or if unusual or unplanned events occur, the program results may be seriously affected and the evaluation report will not accurately reflect the true situation. The monitoring of program activities is called formative process evaluation. The observation of conditions affecting the implementation of program activities is called formative context evaluation.

Formative Process Evaluation

The major purpose for monitoring planned project activities is to identify deficiencies in implementation and to develop strategies for making improvements in the process being followed in time to correct the situation.

Here are some examples of formative process problems:

- Two teachers have decided that the commercial math materials being used in the program are not doing a good job so they have begun substituting their own math.
- Time spent on the teaching of reading varies from classroom to classroom.
- Three instructional aides are teaching reading. Aides were not assigned this responsibility nor have they had training for this task.

Formative Context Evaluation

The major purpose for observing and notating contextual problems that occur throughout the operational period of the program is to be able to plan and introduce alternative actions for alleviating the effects of the problem.

Here are some examples of formative context problems:

- A teacher quits in the middle of the school year.
- The reading resource room is vandalized, and all the equipment and files have been destroyed.
- An epidemic of chicken pox occurs, and over half of the pupils are out of school for a period of four weeks.

Procedures for Collecting Process and Context Data

1. Set up record-keeping forms, such as monitoring forms, and management records. The forms should be comprehensive, simple, and serve as many purposes as possible. (Two sample monitoring forms follow.)
2. Establish data-collection procedures. Decide who will be asked about activities and operations. Decide when the activities should be monitored. A master schedule should be developed.
3. Develop procedures for acting upon problem situations that require change.

School ____________________

Date ____________________

Evaluator ____________________

SAMPLE MONITORING FORM A

PROGRAM ACTIVITY	DOCUMENT AVAILABLE	PERSON RESPONSIBLE FOR IMPLEMENTING ACTIVITY	HELP REQUIRED	FROM WHOM	COMMENTS

School ______________________

Date ______________________

Evaluator ______________________

SAMPLE MONITORING FORM B

PROGRAM ACTIVITIES	IS ACTIVITY TAKING PLACE AS PLANNED?			EVIDENCE AVAILABLE				
	Yes	No	Not Known	Observation	Records	Conference	Other	None
1. ______________________ (List Activities To Be Monitored)	☐	☐	☐	☐	☐	☐	☐	☐
2. ______________________	☐	☐	☐	☐	☐	☐	☐	☐
3. ______________________	☐	☐	☐	☐	☐	☐	☐	☐
4. ______________________	☐	☐	☐	☐	☐	☐	☐	☐
5. ______________________	☐	☐	☐	☐	☐	☐	☐	☐

PROGRAM EVALUATOR'S GUIDE

Section F

ANALYZE EVALUATION DATA

The Evaluation Improvement Program

PRECIS

A good program evaluation design, pertinent measurement instruments to use in executing the design, and careful administration of those instruments to collect the necessary information prepare the way for data analysis. There is a large array of statistics that can be used--some that are descriptive, others that are inferential.

Descriptive statistics include measures of central tendency (mean, median, mode), measures of variability (range, standard deviation, variance), and distributions that are other than normal (skewed, bimodal, rectangular). The inferential statistics presented here include t-tests, analysis of variance, and multiple-regression analysis (which are useful in analyses of test score data), several statistical tests for treatment of ordered and ranked data, and the chi-square test for use with category data in testing frequencies experienced against those expected and to test cross-category associations or the relationship of two variables.

The presentations in this section are on an introductory level. They are not meant to make statisticians of readers of this Guide. They will, however, show what sorts of data analyses are required in those program evaluations that deal with "hard data" and they suggest implicitly what kinds of people might be sought to bring a program evaluation satisfactorily past the data-analysis stage.

CONTENTS

Page

1. INTRODUCTION F-1
    A Working Definition F-1

2. DESCRIPTIVE STATISTICS F-2
    Measures of Central Tendency F-6
    Measures of Variability F-8
    Distributions Other Than Normal F-16

3. INFERENTIAL STATISTICS F-18
    Score Data F-18
    Ordered Data F-19
    Category Data F-19
    Statistical Tests for Score Data F-19
    Statistical Tests for Ordered Data F-22
    Statistical Tests for Category Data F-25
4. DATA INTERPRETATION GUIDELINES F-30

LEARNING EXERCISE 15: COMPUTATION OF X^2 . . . . F-32

1. INTRODUCTION

This section serves as a brief introduction to statistics and data analysis. Because some program evaluators may have received little formal training in this area, we have prepared the exercises, concepts, and examples with this in mind. We do not expect that you will learn all about statistics from this brief treatment. If you complete the section feeling more comfortable with the statistical notions expressed here, feeling that you have an intuitive grasp of the concepts and that you have a better idea of what data analysis can buy in the way of useful information for program evaluation, then we will have achieved our purpose.

A Working Definition

First, it is necessary to have in mind the difference between descriptive and inferential statistics.

Types of Statistics

• Descriptive	Numbers that describe a set of data
• Inferential	Numbers which enable one to test hypotheses and make inferences about the effectiveness of a program

The following demonstrations will illustrate a number of different descriptive statistics:

2. DESCRIPTIVE STATISTICS

Coin-Flipping Demonstration

Directions:

1. First take a penny, or other coin, from your pocket.
2. Flip it 10 times and tally the number of heads and tails in the space below:

Number of Heads and Tails in 10 Flips

Example

Heads	Tails
IIII	~~IIII~~ I

Participant's Individual Tally

Heads	Tails

3. Combine your results with others at your table in the space below:

Example

Participant No.	Heads	Tails
1	4	6
2	7	3
3	5	5
4	2	8
5	6	4
6		
7		
8		
Total	24	26

Participant's Group Record

Participant No.	Heads	Tails
1		
2		
3		
4		
5		
6		
7		
8		
Total		

4. At this point, the workshop trainer will get totals from each table and record them on a master record for the total group.

Theoretically, if a coin is flipped a very large number of times, the number of heads and tails could be expected to be approximately equal. By this time, you have probably noticed that there is random variation from this expected 50/50 split among the individuals sitting at your table and among the different tables in the room. You have probably also noticed that as the number of coin flips increases, the deviation from a 50/50 split becomes less (i.e., your overall table totals come closer to 50/50 than individual totals in your table group; when all tables are combined, the total comes closer to 50/50 than when several are combined).

This concept is basic to what happens with test scores. A single test score of a single student is always an "estimate" of his true score. There are many reasons why we cannot expect to get an exact score, some related to the inherent difficulties in making these kinds of measurements and some related to the specific conditions under which data are collected. However, when groups of students are combined and as the groups become larger, we can have increased confidence that the overall group picture is a better representation of that group's status.

This is also related to the concept of randomness. Those persons who got 1/9, 2/8, or 3/7 splits on flipping their coins deviated randomly from the expected 5/5 split. However, there was a tendency for those to be balanced out by persons who got 9/1, 8/2, or 7/3 splits. This is how random selection of students and random assignment to groups can assure that certain factors not controlled by the design are controlled through random selection or random assignment.

At this point, we have certain basic raw data but they are not yet organized in a very systematic fashion. Now we will organize the data into a frequency distribution, display them graphically, and get some descriptive statistics. Using just the information on number of heads will demonstrate the point.

5. For your future reference, copy in the frequency column the numbers compiled by the trainer for the total group.

Value (No. Heads)	Frequency	Value x Frequency
10		
9		
8		
7		
6		
5		
4		
3		
2		
1		
0		
Total		
Mean ($\bar{x}$)		
Median (Md)		
Mode (Mo)		
Range		

6. Using the frequencies obtained from the total workshop group, draw a line graph below following the instructions given.

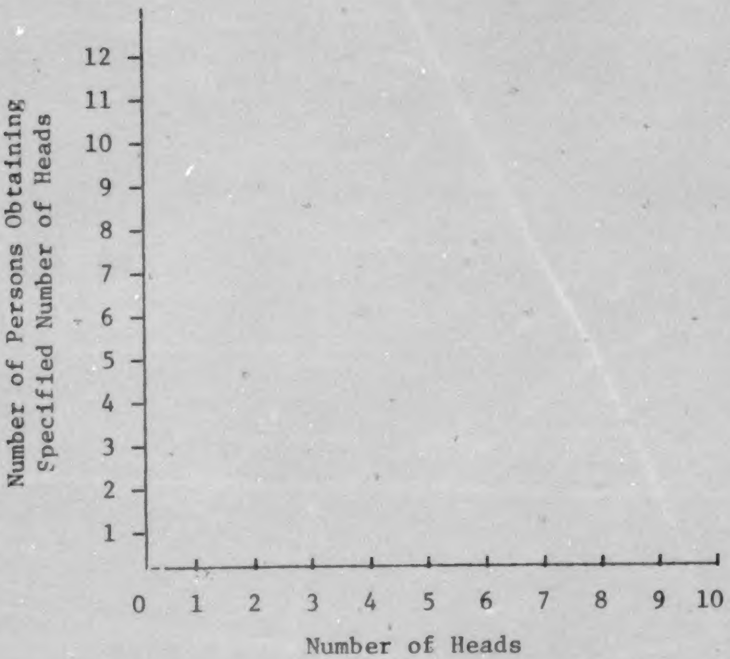
Instructions:

a. What was the highest number of heads anyone had? How many persons had that number?

b. Place an x in the graph opposite these two numbers on the respective scales.

c. Do this for each value and frequency.

d. Connect the x's with a line.

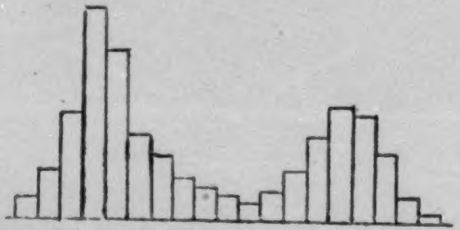


Measures of Central Tendency

There are a number of conventional descriptive statistics which may be used to describe the distribution we just graphed. One is a measure of central tendency. If we select one value which typifies the group, we would select one towards the middle of the range. Three such statistics are commonly used--the mean, median, and mode. The mean is simply the arithmetic average--all the scores divided by the number of persons in the group. In the coin-flipping exercise just completed, to get the mean we could have just listed each person's "score," one after the other, added them, and divided by total number in the group. However, this becomes unwieldy, especially as size of group increases. Instead, we have prepared a frequency distribution. When data are in this form, the score values must be taken into account. To do this, multiply score value by frequency of that score, take the sum and divide by number of persons.

The median is the middle value in a set of scores arranged in ascending or descending order. Count up from the bottom or down from the top. If there is an even number of scores, the median is the average of the two middle scores.

The mode is the most frequently occurring value. Sometimes distribution of test scores will be bimodal--this is often seen in heterogeneous classes where there is a proportionately high number of bilingual students whose command of English is not as good as that of native-speaking English students.



Bimodal

Just a few extremely high or extremely low scores can sometimes affect the mean substantially. For this reason, the median is sometimes preferred. Suppose, for example, in an achievement test, a class of 25 students (Case 1) obtained the following scores (out of 100):

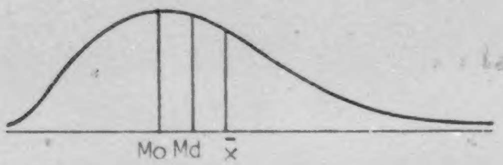
RESULTS OF ACHIEVEMENT TEST

N=25

Case 1			Case 2		
Student No.	Score		Student No.	Score	
1	98		1	98	
2	98		2	98	
3	97		3	97	
4	81		4	81	
5	80		5	80	
6	79		6	79	
7	78		7	78	
8	77		8	77	
9	75		9	75	
10	75		10	75	
11	73		11	73	
12	72	Range = 98-60 = 38	12	72	Range = 98-20 = 78
13	71	Median = 71	13	71	Median = 71
14	71	Mean = 73.6	14	71	Mean = 68.8
15	70		15	70	
16	68		16	68	
17	68	← Mode	17	68	← Mode
18	68		18	68	
19	67		19	67	
20	65		20	65	
21	65		21	65	
22	63		22	63	
23	61		23	21	
24	60		24	20	
25	60		25	20	

The median (in this example, the 13th score) is 71. The mean is 73.6. But now suppose that the lowest three scores are 21, 20, and 20 instead of 61, 60, and 60 (Case 2). The median remains the same, but the mean is now 68.8, a drop of 4.8 points. Sixty percent of the pupils scored 70 or better, but the mean does not reflect that. It has been affected by three uncharacteristically low scores. In this illustration, the median would be a better indicator of central tendency than the mean.

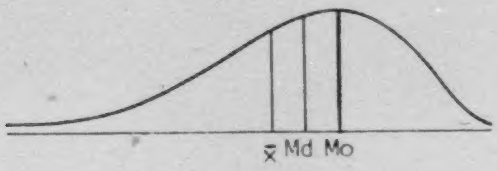
The three measures of central tendency will vary in their relationship to each other depending on the shape of the distribution. The two illustrations below demonstrate this:



$\bar{x}$ = Mean

Md = Median

Mo = Mode



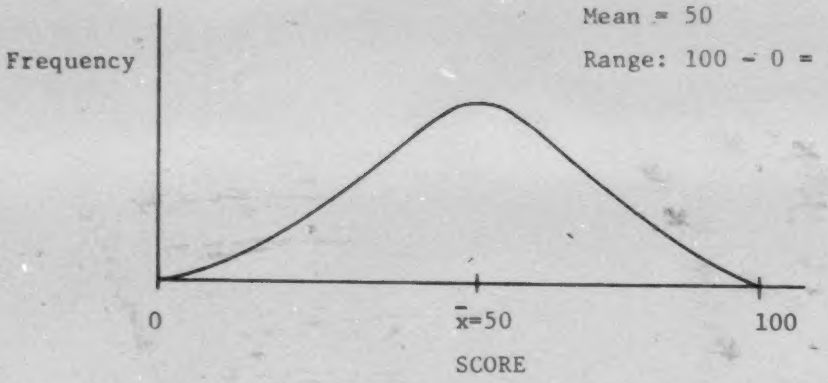
Measures of Variability

Giving the central tendency is necessary but not sufficient to adequately describe a set of data. The amount of variation in scores is also important to consider. Three such statistics will be discussed--the range, standard deviation, and the variance.

The range. In the example we just used to demonstrate the effect of extreme scores on the mean, the range of scores for the first case was 98-60 or 38 and for the second case it was 98-20 or 78.

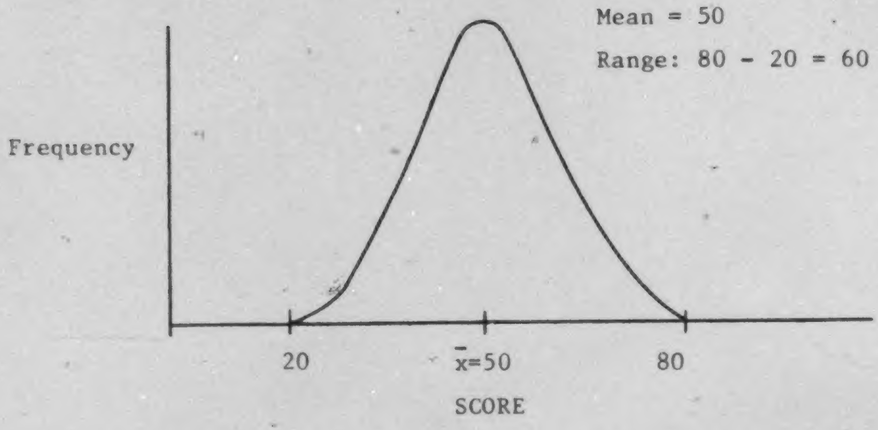
Here are some other examples of measures of variability:

Case 1



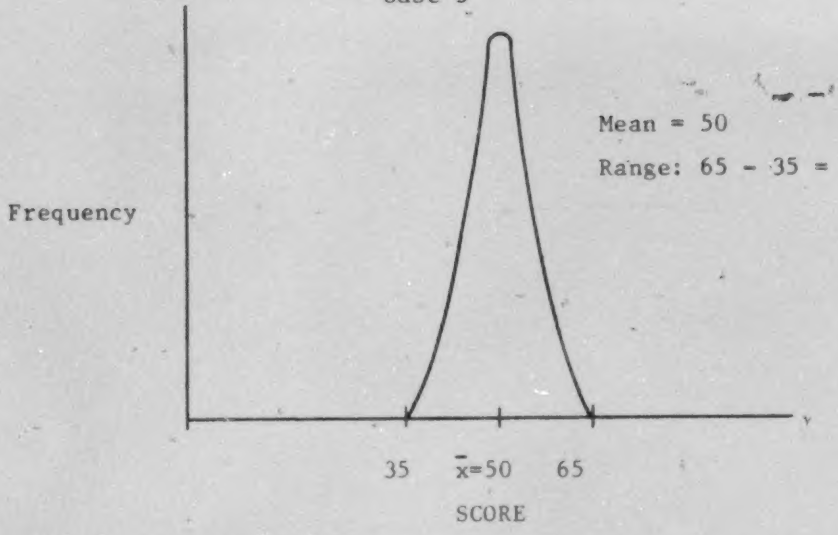
Mean = 50
Range: 100 - 0 = 100

Case 2



Mean = 50
Range: 80 - 20 = 60

Case 3



Mean = 50
Range: 65 - 35 = 30

The standard deviation. Another statistic that will help us as we look at distributions of scores is the standard deviation. This measure tells us more about how the scores spread themselves around the mean or average score. The closer the scores cluster around the mean, the smaller the standard deviation.

Continuing with the same three cases, the figures on page F-11 show how the size of the standard deviation from the mean reflects the spread or variability of scores. The more spread out or variable the scores, the larger the standard deviation.

The familiar bell-shaped curve, or normal distribution, shown on page F-12, forms the basis for making statistical interpretations, and there are known relationships between standard deviation units and the percent of cases falling within those units.

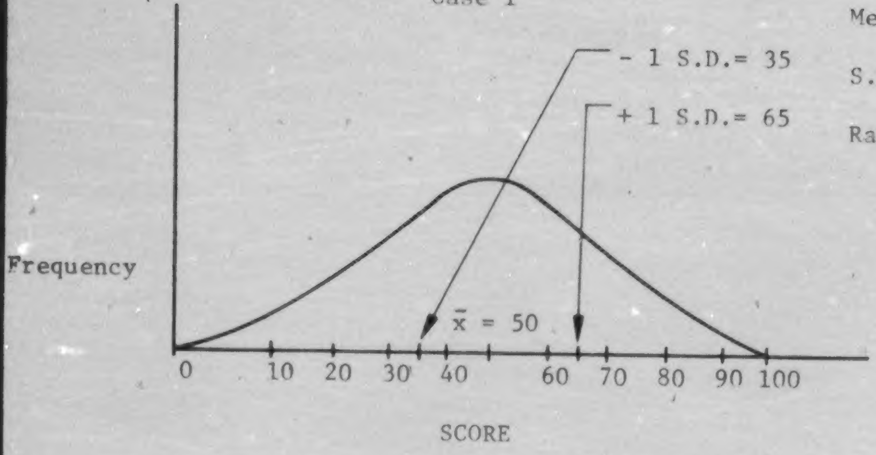
In the theoretical curve, 68 percent of all scores lie between (+) and (-) one standard deviation; 95 percent of the scores lie between (+) and (-) two standard deviations, and almost all lie within (+) and (-) three standard deviations. This relationship enables us to determine the likelihood that differences between two or more groups or two or more sets of scores obtained at different times are significantly different from each other.

The normal distribution does not exist in nature. It is an idealized mathematical distribution which approximates many "real" distributions that are found in nature. Its usefulness lies in the fact that known percents lie within given standard deviation units.

In Case 1, with a mean of 50 and a standard deviation of 15, 68 percent of the scores would fall between 35 (mean - 1 S.D.) and 65 (mean + 1 S.D.). Further, 95 percent of the scores would fall between 20 and 80, while 99 percent of the scores would fall between 5 and 95.

If two individuals in Case 1 made raw scores of 50 and 65, their percentile ranks (P.R.) would be 50 and 65. These same raw scores in Case 3 would yield P.R.s of 50 and 99.

Case 1

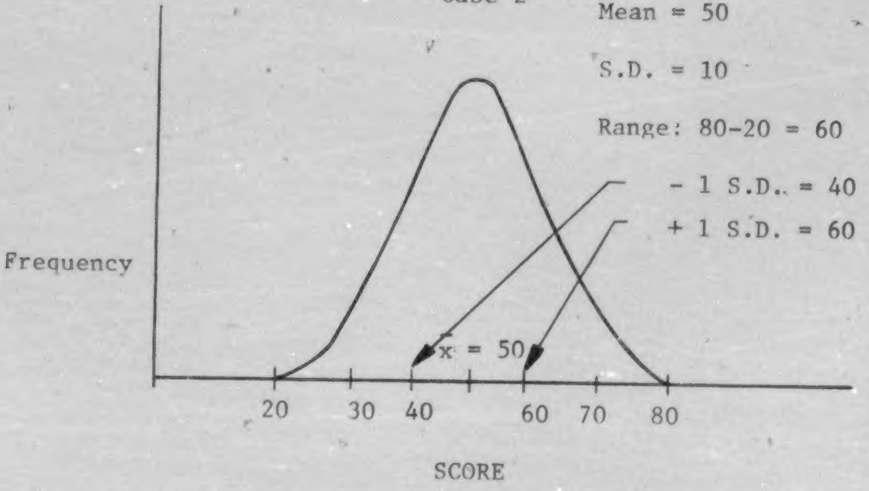


Mean = 50

S.D. = 15

Range: 100 - 0 = 100

Case 2

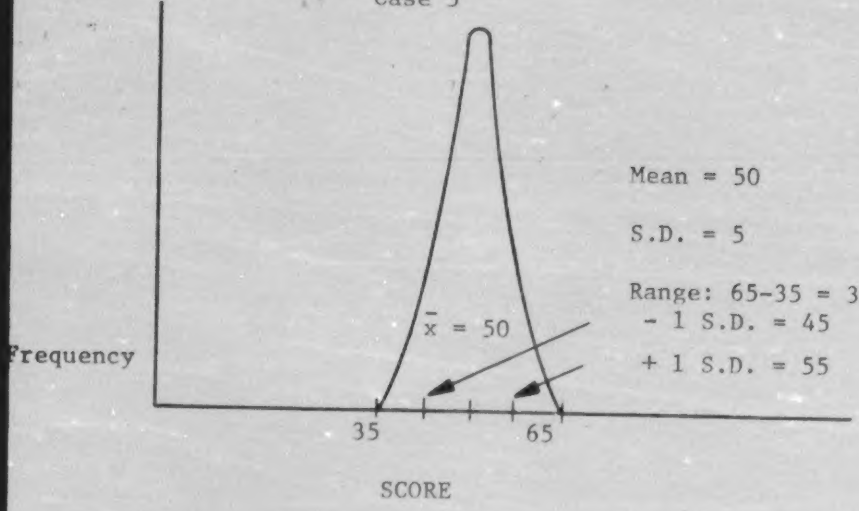


Mean = 50

S.D. = 10

Range: 80-20 = 60

Case 3



Mean = 50

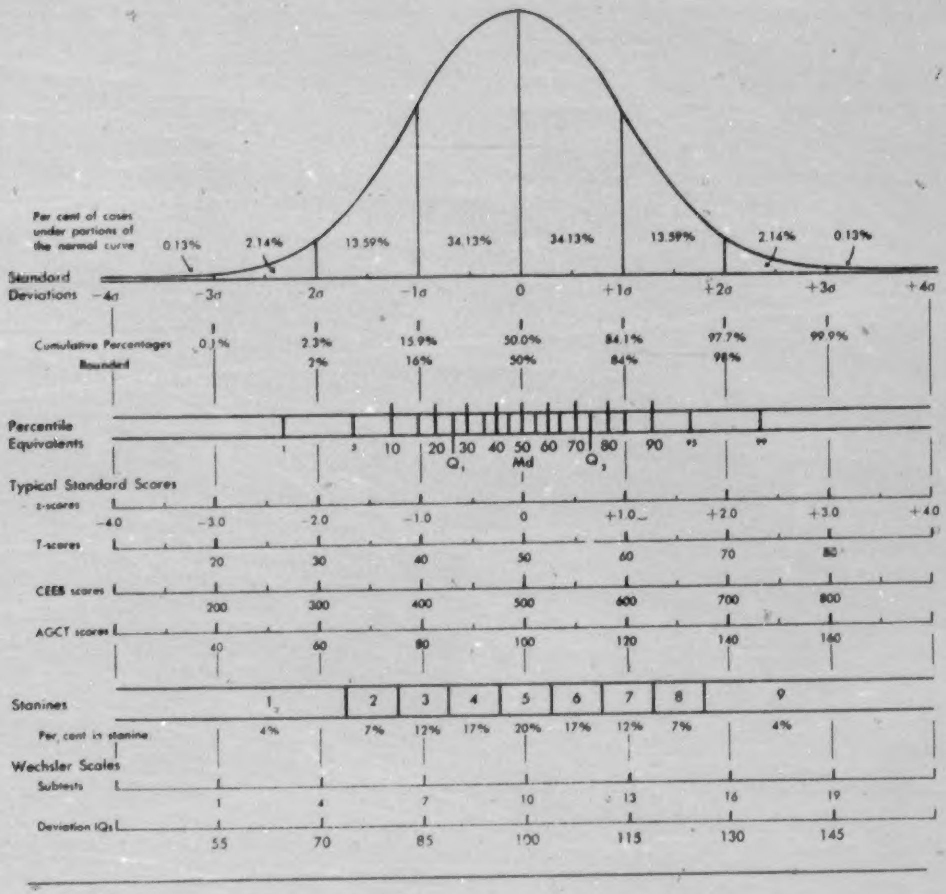
S.D. = 5

Range: 65-35 = 30

- 1 S.D. = 45

+ 1 S.D. = 55

F-12



The Psychological Corporation. Test Service Bulletin No. 48, January 1955. Printed by permission of the publisher, Psychological Corporation.

Note also the relationship among various kinds of scores discussed in the previous section on instruments. In the figure on F-12, the unequal units on the percentile scale can be easily seen. This has important implications for the kinds of statistical tests that can be used. Four different kinds of standard scores are shown--each with a different mean and standard deviation. Stanines are also shown in their relationship to other kinds of scores.

Now, looking at the curves in the three different cases, notice they all have the same mean, median, and mode. But the spread of scores differs markedly across the different cases. So far, we've just talked in the abstract about the spread of the scores. Now let's consider an example.

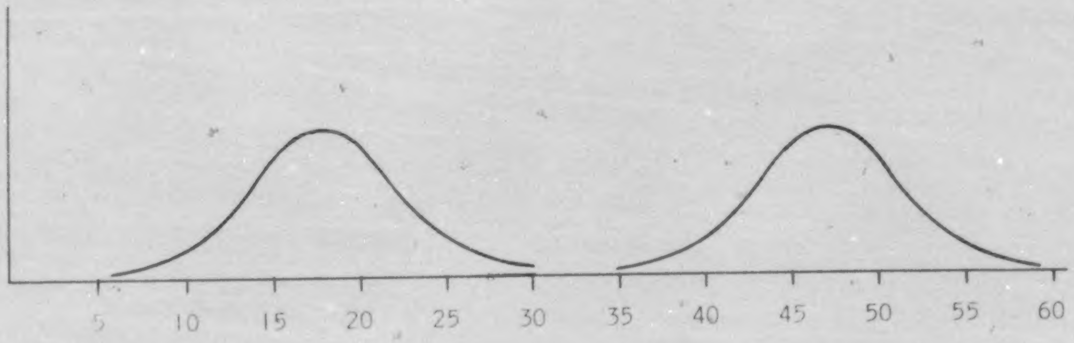
> As an evaluator, you are collecting a variety of measures by which you intend to see to what extent your program has met its objectives. For example, one anticipated outcome is an increase in achievement. To measure this, suppose you administered a standardized achievement test at the beginning and end of the year.
>
> The mean on the pretest was 67; the posttest mean was 70, a mean increase of 3 points. That is a change, but can you now say, "The program was a success. Our kids gained three points on an achievement test administered on a pre-post basis."?

If you answered that question "Yes," you were mistaken. Think again. Remember when we changed the last three scores in our table, "Results of Achievement Test," the mean dropped from 73.6 to 68.8, a drop of 4.8 points? Yet only three scores changed! Now think of the first 25 scores as a pretest and the second set as a posttest. You would think twice about reporting a loss of 4.8 points without some careful examination of the data. So be as skeptical about that increase of 3 points in the above example as you are, rightfully, about that loss shown in Case 2 of the table.

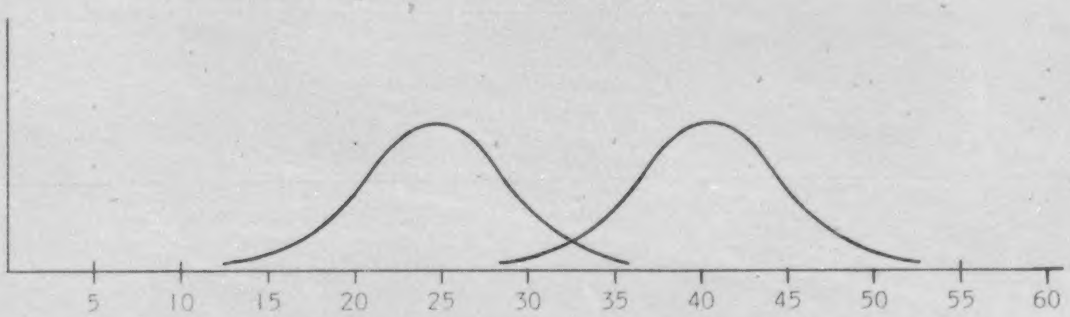
The variance. The variance is closely related to the standard deviation, and mathematically it is simply the square of the standard deviation.

Statistical tests that are made to see if there has been "real" gain between pre- and posttests or those made to see if "real" differences exist between two or more groups are called tests of significance. They examine the difference between means in relation to the variance of the groups and then use the normal curve or other theoretical curves to interpret the results.

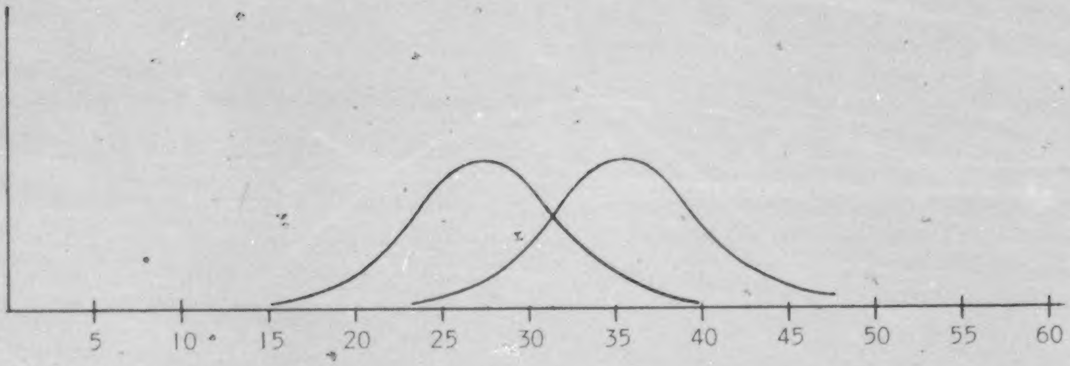
Consider what happens when we want to compare two sets of scores, such as a pretest and a posttest. Ideally we would like to see two nonoverlapping sets of scores:



The lowest score on the posttest is higher than the highest score on the pretest. Clearly, there has been a significant change. Unfortunately, real data do not usually behave this way. Usually there will be less or more overlap in scores, like this:

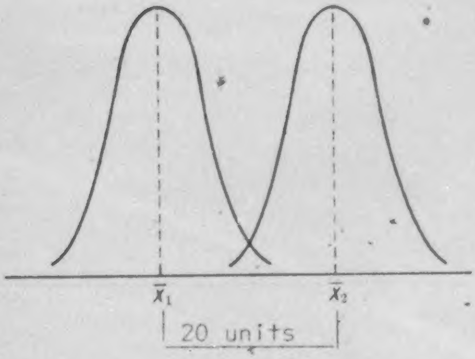


or this:

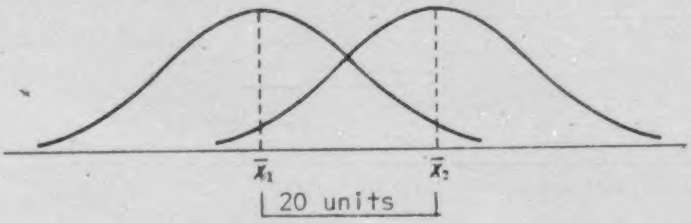


As the overlap between pretest and posttest scores increases, we become less sure that a "real" change has occurred. Inferential statistics looks at the amount of change in relation to the variance and gives the probability that a "real" change has occurred.

The visual comparison below readily illustrates why variance is important.



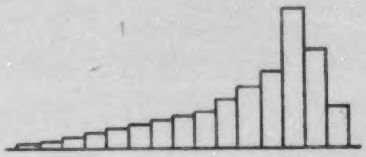
Mean difference of twenty units between two groups with relatively small standard deviations



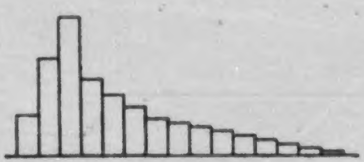
Mean difference of twenty units between two groups with relatively large standard deviations

Distributions Other Than Normal

While test scores from norm-referenced tests generally distribute themselves somewhat like the theoretical normal curves we have been discussing, not all scores do. A basic underlying assumption for many tests of significance is that the data are distributed normally. Before proceeding with any planned analysis, it is a good idea to draw a graph and look at the way the data do distribute themselves. If in doubt, there are procedures for testing whether a given distribution departs too far from normalcy for you to use a statistic based on a normal distribution. Here are some other than normal distributions you may run into:

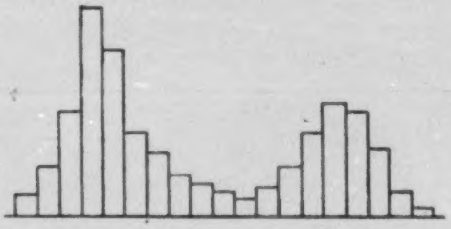


Negatively Skewed

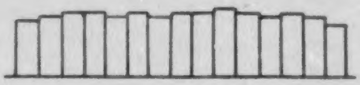


Positively Skewed

If a test is too easy, your scores may look like the figure on the left. There was not enough "ceiling" on the test to adequately differentiate among the best students. If a test is too difficult, the scores may look like the figure on the right with little differentiation among the poorer students. Most norm-referenced tests do not measure well at either extreme of the range for which they are intended.



Bimodal



Rectangular

A bimodal distribution mentioned earlier is one which has scores heavily concentrated in two distinct parts of the scale. This may tend to happen in some bilingual programs if you have a mix of native English-speaking and native Spanish-speaking students neither of which have good command of the other's language. One way to treat this problem is to consider each group separately in the analysis.

A rectangular distribution will be obtained if you plot percentile scores for a group that is similar to the norm group. This is a function of the percentile scale. The kinds of program evaluation questions you want to answer and the kinds of data you use to develop the answers will guide you in the selection of which statistics to use.

In summary, you have been introduced to the following descriptive statistics:

Descriptive Statistics

Frequency	The number of times a given "value" occurs
Mean	The average value; the sum of all values divided by the total number of values
Median	The middle value in a distribution arranged in order from high to low, or from low to high
Mode	The most frequently occurring value
Range	The difference between the highest and lowest value
Standard Deviation	A measure showing how scores spread themselves around the mean
Variance	The square of the standard deviation, the basis for most inferential tests of significance

3. INFERENTIAL STATISTICS

Inferential statistics provide a way to test the significance of results obtained when data are collected. As noted in the discussion on descriptive statistics, all measurement is subject to error (due to inherent difficulties in measuring behavior and to specific testing conditions) and to random fluctuation (due to the particular persons included in the sample). Inferential statistics provide a way to separate chance errors and random fluctuation from real changes.

In selecting a particular statistical test, it is important to know what kinds of data you are dealing with and what the basic assumptions are about those data. Program evaluators will encounter three basic kinds of data:

- score data
- order data
- category data

Score Data

Each person has a numerical score that represents his or her performance or behavior. These are the kinds of data that come from any standardized test.* Techniques used with score data make some rather stringent assumptions:

1. Intervals between scores are equal; that is, differences between scores at one point on the scale are equal to the same size differences at any other point on the scale. (Note: Percentiles and grade equivalents do not qualify on this point.)
2. The scores are assumed to be normally distributed within the population from which they are drawn.
3. The variances in two or more groups being compared are the same.**

---

*Percentile scores excepted

**Actually, the assumption is that the variances in the populations are the same, where population is defined as the total group to which you can generalize.

However, empirical studies have shown that some violation of these assumptions by some tests does not impair their usefulness.

Ordered Data

Each person is assigned a rank that represents his position along a scale. If you have five different texts you are considering for adoption and ask a committee to place them in order from the most preferred to the least preferred, you have ordered data. Sometimes score data can be treated as ordered data, particularly if there is reason to believe the assumptions underlying score data have been badly violated. Statistical tests for ordered data do not make stringent assumptions.

Category Data

Each person is counted as belonging in a particular classification or category. Number of parents for a bond increase election and number of parents against the election constitute category data. Or a comparison which involves numbers in different ethnic groups gives category data.

Statistical Tests for Score Data

If the data you have can meet the assumptions underlying tests for score data, there are many different and potentially powerful tests that can be used. Most inferential tests for score data require special training for their proper selection and use. Unless the program evaluator has had this training, he or she is advised to seek the help of someone who has.

Several commonly used tests will be mentioned, but no effort will be made to teach the computational routines. Program evaluators who have access to a computer center may wish to seek assistance from that source once the decisions have been made as to what kinds of analyses are appropriate. Do not expect computer people to help you decide what analysis is most appropriate. They may be statisticians as well as data processors, but most are not.

The t-tests. A t-test compares two means (pretest vs. posttest or group 1 vs. group 2) to determine whether "real" differences exist. There are several variations in computational routines depending upon the kinds of data being used. In order to select the appropriate computational routine for t-tests, the evaluator must know:

1. Whether the groups being compared are independent or correlated. If you have two measures on each person (pre- and posttest), the groups are correlated. If you are comparing two different groups, use the t-test for uncorrelated means.
2. Sample size. N = 30 for each group being compared is the generally accepted lower limit for using the t-test. For smaller groups, one of the tests for ordered data may be more appropriate. A description of this type of inferential test will be found below.
3. Whether the variances differ markedly. Unless the variances of the two groups are similar, use of the t-test is questionable. A separate test can be made to determine whether unequal variances is a problem.

Note: A t-test used under pre- and posttest conditions must take into account what the expected gain would have been without the special program. Given no special program, average students are expected to gain one month for each month of instruction. To demonstrate superiority of a special program, it should produce gains beyond those expected in the absence of the program. Expectations for an educationally deprived group may be only one-half year for each school year. Past growth history of pupils involved can help determine what this expectation is.

Analysis of variance. In its simplest form, analysis of variance is used when you wish to find out if differences exist in more than two groups. This is a practical method for program evaluators to use.

Analysis of variance can also be used when you wish to examine various factors that may be affecting a program (instructional method, amount of time devoted to instruction, use of teacher aides). This is called factorial design and is potentially a very powerful tool. Unfortunately, its usefulness is somewhat limited by the need to assign either pupils or classes randomly to each possible combination of all variables being investigated. For example:

	Method 1		Method 2	
	30 min.	90 min.	30 min.	45 min.
Aides Present				
Aides Absent				

In this very simple design, you would need to set up 8 different situations (method 1, 30 minutes of instruction, aides present; method 1, 30 minutes of instruction, aides absent; etc.) and then randomly place students into each of the eight situations. Or you could assign intact classrooms (those not divided into subgroups) to each of the eight conditions. But just one classroom per situation is not sufficient to take into account the teacher variable. For this reason, powerful as they are, factorial designs may not be very practical for program evaluation.

Multiple-regression analysis. Multiple-regression analysis deals with prediction. In the case of program evaluation, it might be nice to know which pupils would benefit most from certain instructional units or which combination of program characteristics produces the greatest student achievement.

To set up a multiple-regression analysis, the program evaluator must do the following:

1. Identify a suitable criterion that is acceptable evidence of achievement. (End-of-the-year achievement test may serve very well.)
2. Identify a set of predictors--those things that either preexist or measurements that will be taken during the year that you think will affect student outcome. Preexisting factors may be such things as age, sex, general ability, socioeconomic status, grades in related courses, etc. Predictive measurements taken during the year may be test scores on units of instruction, teacher judgment about pupil progress, pupil self-evaluation, and the like.

Statistical Tests for Ordered Data

Ordered data may be obtained in two basic ways. First: No numerical scores are obtained, but you are able to place persons or objects along some dimension of interest (as when a committee reviews five textbooks up for adoption and can make a series of decisions as to which is most preferred, which is least preferred, and which fit in between). Second (and most common): You have obtained numerical scores but feel the scores are not precise enough to meet the assumptions to use tests for score data. If you must convert score data to ordered data, be aware there are standard conventions for dealing with this:

Scores	Ranks	
13	1	
12	2	
11	3.5	If two scores are equal, the average rank (3 + 4) ÷ 2 is assigned.
11	3.5	
9	5	Rank 5 (not 4) is assigned to next score.

Scores	Ranks	
13	1	
12	2	
11	4	If 3 scores are equal, the average rank (3 + 4 + 5) ÷ 3 is assigned.
11	4	
11	4	
9	6	Rank 6 (not 5) is assigned to next score.

The Sign Test. The Sign Test can be used to determine whether changes have occurred between two different points in time. For example, suppose an evaluator wants to determine the effectiveness of a new unit on citizenship designed to encourage pupils to take a more active interest in a coming community election. The evaluator rates each student on a scale of 1 to 10 before instruction begins by getting information on such things as his or her knowledge of who is running for office, what the issues are, how much time is spent watching local TV newscasts or reading about the election in local newspapers. After the unit, the measures are repeated, and new values on a scale of 1 to 10 are assigned.

Data are recorded as follows:

Pre- and Posttest Scores on Community Election Unit

Pupil No.	Pretest	Posttest	Change[*]
1	2	4	+
2	4	5	+
3	1	3	+
4	3	2	-
5	5	4	-
6	6	8	+
7	1	5	+
8	4	5	+
9	3	7	+
10	2	1	-

Note: Pupils who do not change are eliminated from the table.

The test consists of counting the number of "- changes," noting the total number of students who change in either direction, and consulting a table designed for this test.[**] In this case, the change is not significant.

The Kruskal-Wallis Test. This test can be used to determine whether there are differences among groups.

Suppose the evaluator wants to examine the self concept of students in three different groups (those who have had two years of compensatory education, those who have had one year, and those who have had no exposure to compensatory education). The evaluator gives a self-concept measure and converts the scores to ranks. Data are recorded as follows:

---

[*] + and - are considered a form of ordered data.

[**] For complete description, see Linton, M. & Gallo, P. S. Jr., The practical statistician: simplified handbook of statistics. Monterey, CA.: Brooks Cole, 1975.

Time Exposure to Compensatory Education

2 Years		1 Year		None	
Score	Rank	Score	Rank	Score	Rank
18	3.5	12	1	18	3.5
28	8	16	2	21	5
32	9	37	11	26	6.5
46	13.5	40	12	26	6.5
52	16.5	46	13.5	33	10
62	20	52	16.5	51	15
63	21.5	61	19	53	18
		63	21.5	68	23
				70	24
	$T_1 = 92.0$		$T_2 = 96.5$		$T_3 = 111.5$
	$n_2 = 7$		$n_2 = 8$		$n_3 = 9$

For the computation-minded, your workshop trainer can give you the procedure to follow or consult Linton and Gallo cited on page C-23.

Two other relatively simple tests that can be used with ordered data are as follows:

The Rank Sums Test. This is similar to the Kruskal-Wallis Test, but can be used when there are only two groups to be compared.

The Friedman Test. The Friedman Test is appropriate when more than two measurements are made on the same persons at different times.

Tests for Ordered Data

Test	Use
Sign Test	Tests pre- and postmeasurements on a single group.
Rank Sums Test	Tests differences between two groups.
Kruskal-Wallis Test	Tests differences among three or more groups.
Friedman Test	Tests differences when three or more common measurements are made on the same persons, over time.

Statistical Tests for Category Data

The most commonly used test for category data is the chi-square (X^2) test. However, it may be used in a number of different ways for different purposes. The two most common uses of this statistic for the program evaluator will be 1) to test the deviation of obtained frequencies against some a priori set of expected frequencies, and 2) as a test of association.

Deviation from expected frequencies. To return to our seventh-grade experimental reading program, suppose one of the objectives deals with the attitudes of students in the program. An attitudinal questionnaire is given at the end of the year to see how the group felt about the program. One of the questions the evaluator asks is:

> All things considered, did you enjoy
> the experimental reading program?
>
> The students respond "Yes" or "No."

If the students really have no predisposition toward the program one way or the other, we would expect that about half of them would reply "Yes" and about half of them would reply "No." If the overall response is generally positive, we would expect more than half to reply positively.

Suppose that out of 100 students sampled, 65 students said "Yes" (they enjoyed the experimental reading program) and 35 said "No." Is 65 enough greater than 50 to conclude that the overall response is generally positive and that it did not just occur by chance?

The statistical question is: Is a 65/35 split significantly different from the 50/50 expected by chance if the students really have no predisposition one way or the other? The X^2 test may be used to answer this question. The first step is to construct the table:

Number of Students Who Responded "Yes" and "No"

	Yes	No	Total
Observed Frequency	65	35	100
Expected Frequency	50	50	100

For these data, $\chi^2 = 8.41$. (Those interested in learning how to compute chi-square [χ^2] should see the Learning Exercise that begins on page F-32.)

This result must now be referred to a table to determine whether it may be considered significant. To use the table, it is necessary to know the degrees of freedom for this problem and to select a level of significance. The concept of degrees of freedom is related to the number of categories being treated. For this kind of problem, the number of degrees of freedom is one less than the number of categories. Since there are two categories, there is one degree of freedom.

The selection of level of significance is somewhat arbitrary and indicates the amount of risk the evaluator is willing to take. The greater the magnitude of an observed difference in relation to the variability of the score involved, the more likely it is that a real and significant difference does exist. It is statistically possible to state the chances that an observed difference is a real one or one due to chance. In our example, we will select the .05 level of significance. This means the evaluator is willing to run the risk of being wrong five times in 100 if he assumes that all differences larger than the one read from the table are considered to be real and significant. It is also important to note that sample size becomes important when attempting to establish whether or not there is statistical significance. The larger the number of observations, the greater the opportunity is for the effects of chance to be reduced.

Now look at a portion of the table on significance levels for χ^2.

Portion of Table Showing Significance Levels for χ^2

Degree of Freedom	Significance Levels .25	.10	.05	.025	.01	.005
1	1.3	2.7	3.8	5.0	6.6	7.9
2	2.8	4.6	6.0	7.4	9.2	10.6
3	4.1	6.3	7.8	9.4	11.3	12.8
.	5.4	7.8	9.5	11.7	13.3	18.5
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

This table gives X^2 values for significance levels from .25 to .005. Since we selected the .05 level and we have one degree of freedom, the value of interest to us is 3.8. In order for the difference found in our problem to be considered significant, our X^2 value has to be greater than 3.8. Since our value is 8.41, we can conclude that students in this group really do have a generally favorable attitude toward the program. In fact, our value is greater than 7.9, the value given at the .005 level of confidence. A value as large as 8.41 occurs less than once in 200 times.

X^2 as a test of association. The most frequent use of X^2 is as a test of association. This test will tell you whether or not there is a relationship between two variables. For example, you may survey your community to get information about whether they would support a tax increase to provide additional school services. Because having or not having children in school may influence the vote, you want to analyze your data to see if there is a relationship between the responses you got and having children in school.

Responses on Tax Increase Issue

	Have Children in School	Do Not Have Children in School	Total
Approve Increase in Taxes	60	20	80
Do Not Approve Increase in Taxes	30	40	70
Total	90	60	150

For these data, $X^2 = 14.76$, again a highly significant value (see computation on page F-32). We can conclude that there is a definite relationship between having children in school and willingness to support a tax increase. Course of action: Get the parents out to vote!

χ^2 as a test of association can accommodate more than two levels for each variable, provided certain conditions are met. In the above example, there could have been three categories of response--"for," "against," and "undecided." Or you may have wanted to do the analysis by age group of respondent (21-35, 36-50, over 50), or by some economic index (high, medium, and low), or by ethnic group (white, black, Chicano, other). The degrees of freedom change as the dimensions of the table change and are equal to

$$df = (r-1)(c-1)$$

where r equals number of rows and c equals number of columns. For a 3 x 4 table, $df = (3-1)(4-1) = 6$. So long as the rules shown below are observed, χ^2 can be a very flexible tool.

Rules to Follow When Using χ^2

1. The raw data must always be frequencies. Counting people who pass or fail a test is legitimate. Counting the number of items that each person passes and getting an average score is not legitimate (this is score data). If your data are presented as percentages, convert back to frequencies.
2. All χ^2 analyses require that each subject or event be counted only once. In some cases, you may have more than one measure of a given type on each person. Special techniques must be used when this occurs.
3. If samples are very small, or if some expected events are extremely infrequent, χ^2 may not be appropriate. There must be expected frequencies for 2 x 2 tables of at least 5 tallies in each cell. For larger tables (2 x 3 or greater), all expected frequencies must be 2 or more. Special tests can be applied to make adjustments if this criterion is not met.
4. When something is counted because it is present, absence must also be counted. For example, if you wish to see if sex is related to passing or failing some objective, you must record failures as well as passes in the two groups.

SUMMARY OF INFERENTIAL STATISTICS

Kind of Data	Statistical Tests	Purpose	Example of Question Asked
Score Data	t-Test	To determine whether a significant difference exists between two groups	Did students in the demonstration program perform better on a test of achievement at the end of the year than pupils in the regular program?
		To determine whether a significant difference exists between pretest and posttest	Did a significant change take place over normally expected gain during the course of the year?
	Analysis of Variance	To determine whether significant differences exist among three or more groups	Is student achievement affected by cutting instructional time from 60 minutes to 50 minutes or 40 minutes?
		To determine what factors account for outcomes of a particular program	Are gains in student achievement due primarily to teaching methods, to time allotted for instruction, or to the presence of aides in the classroom?
	Regression Analysis	To predict what factors account for student outcomes	Is it possible to predict which students will benefit most from a unit on alcohol abuse?
Ordered Data	Sign Test	To determine whether a significant change has taken place between two different testing times	Did students take a more active interest in the community election after a special unit on citizenship?
	Rank Sums Test	To determine whether there is a significant difference between two groups	Are students ranked differently on aggressive behavior in school A as compared to school B?
	Kruskal-Wallis Test	To determine whether significant differences exist among three or more groups	Do the self concept of pupils with varying degrees of exposure to compensatory education differ from one another?
	Friedman Test	To determine significant differences when three or more common measurements are made on the same persons over time	Do students' perceptions of their teacher change over the course of a semester?
Category Data	Chi-Square (χ^2)	To test deviation of obtained frequencies against some a priori set of expected frequencies	Did parents make favorable responses significantly more times than could be expected by chance?
		To determine whether there is a significant relationship between two variables	Are parents with children in school more likely to favor a tax increase election than persons not having children in school?

4. DATA INTERPRETATION GUIDELINES

Once the analyses have been performed and certain outcomes have attained statistical significance, and once the descriptive data have been summarized and presented in tabular and graphic form:

> What are you justified in saying about the results of the evaluation? What cautions must be observed? What kinds of remarks avoided?

As a general rule, the evaluator is advised not to make broad, sweeping, global statements that the data "prove" the success of a program. Statistics do not prove anything. Statistics provide the basis upon which people make inferences and interpretations. Be sure you distinguish between the facts given by statistics and the inferences made by people.

Moreover, the evaluator must be careful to define the population to which the results are generalizable, citing sampling techniques used to support claims of generalizability. For example, suppose a questionnaire intended to obtain a random sample of teacher opinions about an innovation drew a response from a disproportionate number of female teachers. The evaluator would have to decide how much stock to place in the questionnaire responses and would have a responsibility to report his or her professional judgment on the possible effect of lacking randomness.

Furthermore, the evaluator needs to know and report the relative strengths and weaknesses of the various instruments used. It is advisable to acknowledge the difference between data collection instruments which require people to perform or demonstrate what they know as opposed to just asking them to make judgments or offer opinions. Judgments, particularly when made about other people, are prone to large fluctuations due to differences which exist among people because of their varying standards and background influences.

Thus, with a good design and appropriate analysis, the evaluator at a minimum should be able to say:

1. Which students or student groups are realizing achievement and other benefits from the program and which are not;
2. Which components of the program are paying off in student gains and improvements, and in what ways;
3. What impacts other than changes in student learning have there been which have affected parents, students, teachers, administrators, and others.

LEARNING EXERCISE 15: COMPUTATION OF χ^2

In 2 x 2 table:

	Have Children in School	Do Not Have Children in School	Total
Approve Increase in Taxes	60 (a)	20 (b)	80 (a + b)
Do not Approve Increase in Taxes	30 (c)	40 (d)	70 (c + d)
Total	90 (a + c)	60 (b + d)	150 (a+b+c+d) = N

$$\chi^2 = \frac{N(|bc - ad| - \frac{N}{2})^2}{(a + b)(c + d)(a + c)(b + d)}$$

$$\chi^2 = \frac{150(|600 - 2400| - 75)^2}{(80)\ (70)\ (90)\ (60)}$$

$$\chi^2 = \frac{150(1725)^2}{30,240,000}$$

$$\chi^2 = \frac{446,343,750}{30,240,000} = 14.76$$

In tables larger than 2 x 2:

The computational scheme for tables larger than 2 x 2 requires that an expected frequency be developed for each cell in the table. The expected cell frequency is obtained by multiplying the total of the row to which the cell belongs by the total of the column to which the cell belongs and then dividing by the grand total.

In this example, data are arranged for an analysis of the returns from a questionnaire which asked parents of three different ethnic groups how many pupils in their school needed a bilingual program.

	Ethnic Group of Parents						
Students Needing Programs	I		II		III		Total
All	75	(47.1)	54	(68.2)	12	(25.7)	141
Most	64	(66.1)	106	(95.8)	28	(36.0)	198
Few	28	(53.8)	82	(78.0)	51	(29.3)	161
	167		242		91		500

The expected frequencies are given in parentheses ( ) and the 47.1 given in the first box is

$$\frac{(141)\ (167)}{500} = 47.1$$

χ^2 is calculated by subtracting the expected value from the obtained value, squaring and dividing by the expected value. When this has been done for each cell, the results are added.

$$\chi^2 = \Sigma \frac{(fo - fe)^2}{fe}$$

$$\chi^2 = \frac{(75 - 47.1)^2}{47.1} + \frac{(54 - 68.2)^2}{68.2} + \frac{(12 - 25.7)^2}{25.7}$$

$$+ \frac{(64 - 66.1)^2}{66.1} + \frac{(106 - 95.8)^2}{95.8} + \frac{(28 - 36.0)^2}{36.0}$$

$$+ \frac{(28 - 53.8)^2}{53.8} + \frac{(82 - 78.0)^2}{78.0} + \frac{(51 - 29.3)^2}{29.3}$$

$$= 58.38$$

χ^2 Exercises

Directions:

1. Using the 2 x 2 method of computing χ^2 as a test of association just illustrated, compute χ^2 for these values, read the level of significance from the χ^2 table at F-26, and draw a conclusion about these data.

	Have Children	Do Not Have Children	Total
Approve	50	30	80
Do Not Approve	40	30	70
	90	60	150

$$\chi^2 = \frac{N(|bc - ad| - \frac{N}{2})^2}{(a + b)(c + d)(a + c)(b + d)}$$

2. Assume your data have a third category of response--"undecided." Compute χ^2 for this 2 x 3 table using the formula

$$\chi^2 = \Sigma \frac{(f_o - f_e)^2}{f_e}$$

determine the level of significance, and draw a conclusion about these data.

	Have Children	Do Not Have Children	Total
Approve	50	20	70
Undecided	10	10	20
Disapprove	30	30	60
Total	90	60	150

ANSWERS

1.

$$\chi^2 = \frac{(|1200 - 1500| - 75)^2}{30,240,000}$$

$$\chi^2 = \frac{150(225)^2}{30,240,000}$$

$$\chi^2 = \frac{7,593,750}{30,240,000} = .25$$

Sig. > .25 (There are more than 25 chances in 100 that the observed differences are due to random fluctuations.)

Conclusion: There really aren't any differences of opinion between persons who have children in school and those who don't.

2.

$$\chi^2 = \Sigma \frac{(f_o - f_2)^2}{f_e}$$

	Have Children	Do Not Have Children	Total
Approve	50 (42)	20 (28)	70
Undecided	10 (12)	10 (8)	20
Disapprove	30 (36)	30 (24)	60
Total	90	60	150

$$\chi^2 = \frac{(50 - 42)^2}{42} + \frac{(20 - 28)^2}{28} + \frac{(10 - 12)^2}{12} + \frac{(10 - 8)^2}{8} + \frac{(30 - 36)^2}{36} + \frac{(30 - 24)^2}{24}$$

$$\chi^2 = 1.52 + 2.29 + .33 + .50 + 1.00 + 1.50$$

$\chi^2 = 7.14$ df = (2 - 1)(3 - 1) = 2

Sig. at .05 level (There are 5 chances in 100 that the observed differences are due to random fluctuations.)

Conclusion: There is a relationship between having children in school and the opinion adults held.

PROGRAM EVALUATOR'S GUIDE

Section G

REPORT EVALUATION RESULTS

The Evaluation Improvement Program

PRECIS

Program evaluation reporting is largely a matter of good school-community relations. The principles that apply to positive and open school public relations apply here as well. Some of these principles, particularly those that are significant in the framework of interim and annual program evaluation and the impact of both on longer-term improvement, are treated briefly in this section.

Relevance, clarity, and specificity are the three critical characteristics of the program evaluation report. It should address each of the program's objectives and report forthrightly on whether or not the data indicate that such objectives have or have not been met. Wording should be clear and concise with modifications of style and approach wherever appropriate to fit various audiences. Statements should be specific enough so that readers will understand what aspects of a program can remain unchanged, what needs changing, and what needs to be quietly laid to rest.

The report should be sent to those who will lend vigorous support to a) the continuation of those parts of the program that have been shown to be successful, and b) improvement of whatever aspects of the program have been shown to be negative or neutral.

CONTENTS

Page

1. IDENTIFYING AUDIENCES G-1

2. WHEN TO REPORT G-1

3. THE INTERIM REPORT G-2

4. PROGRAM MANAGEMENT REVIEW RECORD G-2

LEARNING EXERCISE 16: RECIPIENTS AND USES OF INTERIM EVALUATION DATA G-4

5. THE FINAL PROJECT REPORT G-6

A Suggested Outline G-6

6. REVIEW AND RELEASE OF THE FINAL REPORT G-8

LEARNING EXERCISE 17: DETERMINING APPROPRIATE DATA DISPLAYS G-9

LEARNING EXERCISE 18: WRITING RECOMMENDATIONS FOR THE FINAL REPORT G-14

LEARNING EXERCISE 19: ANALYZING PROGRAM EVALUATION RECOMMENDATIONS G-18

1. IDENTIFYING AUDIENCES

To assure continued program support, it is wise to submit evaluation information to as many audiences as possible. These audiences may include:

- Instructional staff
- Administrative staff
- Parents
- Students
- Citizens' Advisory Committees
- Superintendent
- Board of Education
- Total community
- Funding agency

Agencies vary in the needs they have for evaluation data. The needs, in general, should correspond to the purposes of the evaluation process. The purposes of program evaluation may fall into any of a number of categories, among them the following:

1. Ascertaining program quality for all concerned
2. Providing information for decision makers
3. Improving existing programs
4. Providing satisfaction to participants
5. Communicating with the public

2. WHEN TO REPORT

In planning the program evaluation, the evaluator must determine types of evidence acceptable to each audience. It must also be determined when each audience needs to receive the results of the program evaluation. Some audiences need evaluation reports while the program is in progress. Such reviews are called interim evaluation reports. Other audiences need evaluation reports only at the end of the program in what is commonly known as the final project report. A number of audiences will require both interim and final evaluation reports.

3. THE INTERIM REPORT

The purposes of the interim report are to monitor the program in progress, to derive information that may improve the program, and to get any early indications about the probable outcome. Interim evaluation reporting may be done formally or informally and occasionally, orally. The report should be timely and provide the information needed by specific individuals and groups when they need to act on it. The report should be brief and concise without being cursory, and it should make very clear how the information it contains is to be used. Learning Exercise 16 focuses on the variety of uses and audiences for interim reports. See page G-4.

4. PROGRAM MANAGEMENT REVIEW RECORD

The Program Management Review Record shown on page G-3 can be used both to monitor an ongoing program and to prepare interim reports.

Both program objectives and activities are listed on the form. Accompanying columns allow for recording information on interim progress, specifying the additional assistance that may be needed to sharpen the objectives and facilitate the activities, noting whatever corrective action needs to be taken.

This is a sample of a management support tool that is easy to use, that assists the program evaluator in monitoring the completion of activities, and that provides information for interim reporting so that decision makers may more effectively direct the program. Changes may be made to correct a possibly serious deficiency, or not so critical but nevertheless important omission, in time to make an impact on the final outcomes.

PROGRAM MANAGEMENT REVIEW RECORD

OBJECTIVES AND ACTIVITIES	COMPLETION DATE	COMPLETED		REASON FOR DEFICIENCY (if applicable)	SUGGESTED ACTION TO CORRECT DEFICIENCIES		
		Yes	No		Person Responsible	Action To Be Taken	Completion Date
1.0 OBJECTIVE: By June pupils will have mastered an average of 10 or more comprehension skills as measured by attainment of 80 percent or higher on the criterion-referenced tests accompanying the skills sequence.	June 1974	X					
ACTIVITIES: 1.1 Administer diagnostic test	Sept. 18	X					
1.2 Develop pupil and class profile	Sept. 25	X					
1.3 Place pupils in instructional sequence	Oct. 2	X					
1.4 Establish learning centers	Oct. 16	X					
1.5 Develop independent activities	Nov. 16		X	insufficient time	classroom teacher	assistance by resource teacher	Dec. 1

LEARNING EXERCISE 16: RECEPIENTS AND USES OF INTERIM EVALUATION DATA

This exercise is designed to provide experience working in small groups to determine who needs interim evaluation data and how the information may be used. Three statements of objectives are shown together with a listing of interim information available on each. You are asked to complete the exercise by predicting which groups will need to have each cluster of information and what uses they will likely make of it.

Complete the blanks in column three of the table on page G-11 by indicating for each information cluster one or a combination of the following:

Students

Teachers

Principal

Citizens' Advisory Council

Add others as you wish.

Next, complete the blanks in column four. Sample statements are as follows:

To designate the skills to develop
in the next in-service sessions

To determine whether the objectives
are being met

To determine methods of increasing
parental involvement

INTERIM EVALUATION DATA

School ____________________

Date ____________________

		COLUMN 3	COLUMN 4
OBJECTIVE	INTERIM INFORMATION AVAILABLE	PERSON(S) NEEDING INFORMATION	USE OF THE INFORMATION
By June, 75 percent of the participating pupils will have mastered 10 or more criterion objectives relating to reading comprehension skills. (Check off on profile when teacher determines that the skill has been mastered.)	Number of skills mastered by each pupil		
	Number of students mastering skills A, B, C (etc.)		
By June, at least 40 parents of participating pupils will have provided volunteer help in the classroom, as shown on records kept in the office.	Number of parents involved to date		
	Names of parents not yet involved		
Three-fourths of the staff-development sessions held during the year will be rated as effective by at least 75 percent of the participants responding to a locally developed rating form.	Number of participants rating sessions as effective		
	Rating forms with suggested changes		

5. THE FINAL PROJECT REPORT

The purposes of the Final Report are to summarize the results of the evaluation: What was the program designed to accomplish? What was done to accomplish the objectives? What did the program accomplish? How was the program evaluated? What recommendations are there for further action? Like the interim report, the final project report should be timely, provide the information needed by specific individuals when it is needed, be clear, brief, and concise.

End-of-the-year program evaluation reporting typically is more formal in nature than interim reporting and generally is in written form. One must consider the variety of audiences to whom the final report is to be directed and select the formats, presentations, and visual aids that will be appropriate for each specific group.

The evaluation will convey the same basic information to all audiences; however, the details in the several reports will vary according to the needs and purposes of the several readerships. Whatever the expected readership, brevity and clarity always are paramount considerations.

A Suggested Outline

Below are some suggested headings and guides for writing each section.

1. Program Goals and Objectives
    a. Review and translate the goals and objectives of the program into the language of the reader.
2. Program Description
    a. Describe the population participating in the program. Include the number of pupils, teaching staff, grade level, subject matter, and schools in the study.
    b. State the length of the program with beginning and ending dates.
    c. Describe the significant activities, materials, and personnel used in the program.
    d. Note parts of the program that are unique.

3. Program Evaluation Procedures

    a. Describe the design, instruments, and analyses which were used in evaluating the extent to which the stated objectives were accomplished.

    b. Tailor the language and terminology to the audience that is to receive the report.

4. Program Accomplishments

    a. Describe the positive results of successful activities.

    b. Describe the marginal results of unsuccessful activities.

    c. Describe unanticipated outcomes and side effects that have been observed.

    d. Emphasize changes observed such as score gains, changes in attitudes and behaviors.

5. Program Evaluation Conclusions

    a. Present judgments as to why each objective was or was not met.

    b. Present alternative proposals for different approaches in those instances in which objectives were not realized.

    c. Present alternative proposals for improvements in those instances in which realized objectives could be surpassed in future programs.

    d. Draw summary statements on program effectiveness through a balanced review of successful and not-so-successful outcomes.

    e. Whenever possible, relate program effectiveness to program costs.

6. Other Findings

    a. Report on the results of surveys, questionnaires, interviews, and other such data that may not fall under the heading of Program Accomplishments, but are relevant to program outcomes.

    b. Report on informal findings and conclusions drawn from information assembled outside the framework of the program evaluation.

7. Recommendations Related to the Program and Program Evaluation

   a. Recommend a preferred alternative for each new approach and improvement in the program which would lead to greater achievement of objectives in the future.

   b. Suggest revisions in objectives and in affected program features, especially regarding those objectives that were not met.

   c. Suggest revisions in program evaluation design, instruments, analyses, and procedures that can be applied to subsequent program evaluation efforts.

6. REVIEW AND RELEASE OF THE FINAL REPORT

The evaluator should arrange to have the information in the final report reviewed by selected members of the program staff and by a sample of those for whom it is intended. This review should take place while the report is being written. The reviewers should be asked to verify the description of the program and that the types of information presented are those that are needed, that the formats, explanations, and visual aids in the report are clear, and that the recommendations are appropriate and consistent with existing policies, directives, and guidelines.

The final draft should be reviewed by the project director, chief administrator, and staff who were involved in the collecting and summarizing of information. This will provide a final check on the report's accuracy and appropriateness as well as assurance that the report will have the support of all the program participants.

The publication and release of the final report of a program evaluation is usually the responsibility of the chief administrator.

LEARNING EXERCISE 17: DETERMINING APPROPRIATE DATA DISPLAYS

Examine the documents on the next four pages. Each format contains the same information on pupil reading achievement, but the information is reported in three different displays. You are asked to list in the upper right of each display the audiences in your district who could make good use of information of this type. Judge the effectiveness of each display for the chosen groups, decide whether or not it would be satisfactory as is and what modifications would make the display clearer.

Audiences may include program staff, citizens' advisory council, superintendent, board of education, a special interest group, and the total community. Add others if you wish.

Audiences:

____

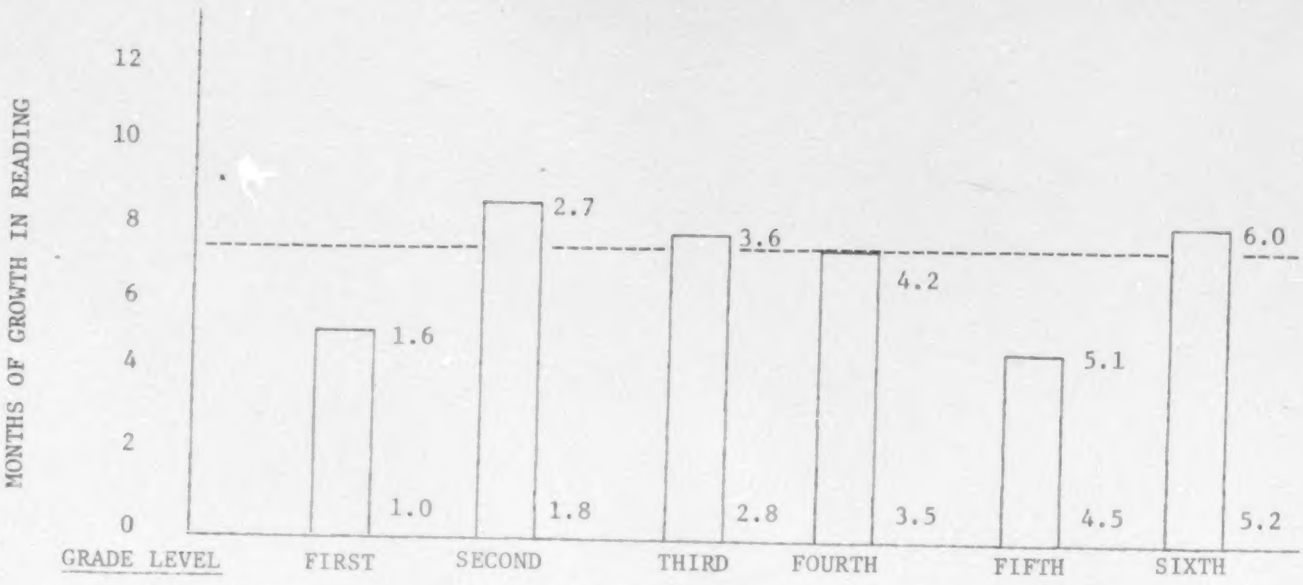
____

____

____

DISPLAY 1

PUPILS' GROWTH ON READING TESTS FROM BEGINNING TO END OF YEAR



The pre- and posttesting covered an instructional period of seven months; therefore, the expected gain is 7.0 months. For example, the mean score on the third grade pretest was 2.8 (eighth month of second grade) and the mean score on the posttest given the same year was 3.6 (sixth month of third grade) for a gain of eight months based on an instructional program of seven months.

DISPLAY 2

SUMMARY OF TESTING FROM THE READING DEVELOPMENT COMPONENT

Grade level (1)	Name of Test (2)	(For State use only) (3)	Form (4)	Level (5)	Months between pre- and posttests (6)	Number of pupils receiving both pre- and posttests (7)	Test results expressed as median grade equivalents			Test results expressed as mean scale scores		
							Pretest (8)	Posttest (9)	Difference (col. 9 minus col. 8) (10)	Pretest (11)	Posttest (12)	Difference (col. 12 minus col. 11) (13)
K												
1	Cooperative Primary Reading		Pre B Pst A	12 12	7	433	1.0	1.6	.6	132	135	3
2	Cooperative Primary Reading		B A	12 23	7	430	1.8	2.7	.9	136	142	6
3	Cooperative Primary Reading		B B	23 23	7	540	2.8	3.6	.8	142	148	6
4	SAT, Total Reading		W	Int 1	7	536	3.5	4.2	.7			
5	CTBS, Total Reading		Pre Q Pst Q	20 30	7	525	4.6	5.1	.5	355	386	31
6	CTBS, Total Reading		Q	2	7	521	5.2	6.0	.8	396	417	21
7												
8												
9												
10												
11												
12												

Learning Exercise 17
G-11

AUDIENCES:

Audiences: ____________________
____
____
____

DISPLAY 3

The objective of this reading program was to increase pupil reading gain at least one month for each month of reading instruction. The program began in September. There was a full teaching staff as well as a complete complement of teacher aides assisting with the program.

The pretests were administered on October 15 and the posttests on May 15. Pupils received seven months of instruction during this program. The tests used included the Cooperative Primary Reading Test Form A and Form B, the Stanford Achievement Test, Form W, and the California Test of Basic Skills Form Q.

Scores achieved on the administered standardized achievement tests varied somewhat between the grades tested. The scores are reported below:

Grade 1	Pupils in grade 1 made six months' growth between the pre- and posttest. The pretest score was 1.0, and the posttest score 1.6.
Grade 2	Pupils in grade 2 made nine months' growth between the pre- and posttest. The pretest scores were 1.8, and the posttest scores 2.7.
Grade 3	Pupils in grade 3 made eight months' growth between the pre- and posttest. The pretest scores were 2.8, and the posttest scores 3.6.
Grade 4	Pupils in grade 4 made seven months' growth between pre- and posttest. The pretest score was 3.5, and the posttest score 4.2.

Grade 5	Pupils in grade 5 made six months' growth between the pre- and posttest. The pretest score was 4.5, and the posttest score 5.1.
Grade 6	Pupils in grade 6 made eight months' growth between pre- and posttest. The pretest score was 5.2, and the posttest score 6.0.

The objective was reached at grades 2, 3, 4, and 6 but was not met at grades 1 and 5.

The objective was exceeded by one month at grades 3 and 6, and exceeded by two months at grade 2.

LEARNING EXERCISE 18: WRITING RECOMMENDATIONS FOR THE FINAL REPORT

This exercise concerns pupils in a program who have had seven months of instruction using a diagnostic/prescriptive teaching approach. Your group will be asked to complete a staff review of the information provided and develop recommendations to be considered by the appropriate decision makers. You are asked to assume that this report is being submitted by a program evaluator to a person in your district who will take some decisive action based on his or her recommendations. In some situations, the recipient would be the principal: in others, the program manager, the superintendent, or the assistant superintendent.

You are asked to write the recommendations section of a final report on the basis of the information in the sections that are included below and on pages G-15 and G-16. Considering this information, what recommendations would you make to the decision maker? What should be left as is? What changes should be made?

EXCERPTS
FROM A FINAL REPORT

PROGRAM OBJECTIVE

By June 1975, the median score for program participants will have increased by one month for each month of instruction as measured by pre- and posttesting on a standardized reading achievement test.

If a class has a median score of 4.3 on November 1 on a standardized reading test and a median score of 5.1 on May 1, the class has gained .8 years or 8 months. Since the instructional time span was 7 months, the objective of one month of gain for each month of instruction has been exceeded.

PROGRAM DESCRIPTION

In part, teachers used the Diagnostic/Prescriptive Teaching (DPT) approach to instruction developed by the Title I Program. This approach involves testing each pupil to determine the reading skills he needs to master during the year. The teacher then uses special materials designed to help each pupil in the area of greatest need. To determine whether instruction was effective, the teacher next assesses the pupil for mastery of those skills. If the pupil has mastered the skill in question, the teacher moves on to work with the pupil in his or her next area of need.

Using this approach, pupils spend less time working on tasks that are too easy or too difficult and thereby spend more time experiencing success with reading tasks at their own respective learning levels.

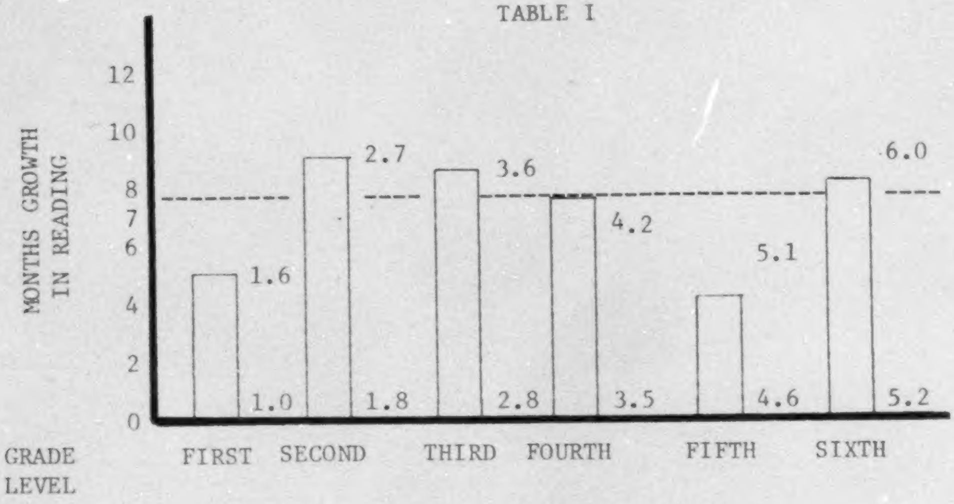
To make this approach work, teachers can and should use a wide variety of instructional materials to help the pupil master needed reading skills.

PROGRAM ACCOMPLISHMENTS

Table I below shows how well the participating pupils did this year in improving their reading skills. Pupils in grade 2, for example, began the year with an average reading score of 1.8. This means they scored the same as most first graders who are in the eighth month of school.

PUPILS' GAIN IN READING ACHIEVEMENT

TABLE I



By the end of the year, second graders were scoring at about 2.7 (an increase of .9). This means they gained nine months in reading skills during the year.

The table shows that pupils in grades 2, 3, 4, and 6 gained at least seven months in reading skills. Pupils in grades 1 and 5 grew six months and pupils in grade 5 grew five months in reading skills.

In summary, the objective was met at four of six grade levels, nearly met at two.

OTHER FINDINGS

A questionnaire was administered to participating teachers. The following findings came from this questionnaire:

- Eighty percent of the teachers reported that more individualized attention could have been given to each pupil if the teachers had received more adult assistance in the classroom.
- Ninety percent of the teachers reported that their pupils had a wide range of academic weaknesses and that it was impossbile to provide adequate help to each pupil.
- Sixty-five percent of the teachers requested additional in-service training in managing the classroom and in grouping pupils for individualized instruction.

CONCLUSIONS

The Diagnostic/Prescriptive Teaching Approach to instruction met the objectives as planned in grades 2, 3, 4, and 6.

First and fifth grade pupils achieved six months' growth during the seven months of instruction. This lower-than-anticipated rate of growth suggests the possibility of problems in the instructional program at these levels that need to be rectified.

RECOMMENDATIONS

Please develop at least three recommendations. After writing them on this page, transfer them to the transparency provided for you. All transparencies will be collected at the conclusion of the exercise and used later in group discussion.

LEARNING EXERCISE 19: ANALYZING PROGRAM EVALUATION RECOMMENDATIONS

Every program evaluation report should contain conclusions and recommendations. These usually are drafted by those who are responsible for analyzing and reporting the data and reviewed by the project director and perhaps by others. Both conclusions and recommendations are based on the information developed during the evaluation process, and on the analyses and interpretations made using that information.

"Rating of Recommendations" sheet which follows contains ten recommendations which have been submitted as parts of a variety of program evaluation reports. In the left margin, the recommendations are consecutively numbered and recorded. In the two columns to the right are spaces to rate each of the recommendations according to two criteria:

Clarity - The wording is clear; you understand what the evaluator is trying to say.

Specificity - The content is specific enough so you have definite clues as to what needs to be done.

Learning Exercise 19

RATING OF RECOMMENDATIONS SHEET

Rate each of the recommendations at the left according to its clarity and specificity. Use a scale from 1 - 3; a 1 means it is clear, a 2 means it is not as clear as it should be, and a 3 means it is not clear. Discuss with others at your table the reasons why you gave any 2 or 3 ratings.

RECOMMENDATIONS	CLARITY	SPECIFICITY
1. Continued emphasis should be placed on individual and small-group instruction.		
2. Decision making relative to the Title I (Compensatory Education) Program should be done whenever possible by those directly participating.		
3. There should be continual evaluation of the elements in the school which can cause or encourage hostility among pupils. Means to eliminate those elements should be developed as soon as possible.		
4. Since parent participation is limited by employment, all activities must be action-oriented and relevant to the pupil's education program.		
5. More emphasis in staff development should be placed on staff attitudes toward the pupil regardless of the pupil's academic achievement.		
6. Before the beginning of the school year, a schedule should be developed for the administration of all evaluation instruments. Regularly scheduled dates should be set for the evaluator to observe project activities to establish the reliability of observational protocols.		

RATING OF RECOMMENDATIONS SHEET (cont'd)

Rate each of the recommendations at the left according to its clarity and specificity. Use a scale from 1 - 3; a 1 means it is clear, a 2 means it is not as clear as it should be, and a 3 means it is not clear. Discuss with others at your table the reasons why you gave any 2 or 3 ratings.

RECOMMENDATIONS	CLARITY	SPECIFICITY
7. Parents and teachers should be actively involved in the evaluation process by knowing the purpose of each instrument and the results as they become available. They should see the evaluation process as a benefit to them in understanding the pupils and how the project can continually be improved by a cooperative effort of staff and parents.		
8. Parent workshops should be given which stress the practical activities involved in conducting a classroom lesson. Material preparation skills both for the classroom and the home should be taught in a practical fashion where parents actively prepare a variety of materials they can use.		
9. Plans for lessons that parents are expected to participate in should be distributed one week in advance to allow the parents time to prepare for the activities.		
10. Better communication systems should be developed to insure that all parents are informed of parent meetings and other activities of the project.		

PROGRAM EVALUATOR'S GUIDE

Section H

APPLY EVALUATION FINDINGS

The Evaluation Improvement Program

CONTENTS

Page

PREPARING TO MAKE MAXIMUM USE OF EVALUATION RESULTS H-1

LEARNING EXERCISE 20: USE OF EVALUATION INFORMATION. H-2

LEARNING EXERCISE 21: ROADBLOCKS TO PROGRAM EVALUATION H-12

PREPARING TO MAKE MAXIMUM USE OF EVALUATION RESULTS

A basic tenet of this Guide is that program evaluation is something that is done with specific purposes in mind, and that evaluation is useless unless those purposes are served. In Section A on purposes and requirements, a number of different purposes were listed and several possible audiences identified for whom the respective purposes seem appropriate. It was suggested that the different audiences likely would have quite different purposes needing to be served through a program evaluation and therefore would want different kinds of information to meet their needs. In Section G, on reporting, diversities of purpose and audience and the consequent needs for tailoring program evaluation components to meet those diversified requirements were again emphasized. The steps outlined in those two sections are probably the most productive things an evaluator possibly can do to ensure that effective use will be made of the evaluation findings, conclusions, and recommendations. Brief summaries of these steps follow:

1. Determine all the purposes the program evaluation is to serve.
2. Make explicit various questions that all users would like to have answered in satisfying their program evaluation needs.
3. Identify the kinds of information that will prove acceptable as evidence bearing upon those questions.
4. Provide interim reports during the progress of the program to give early evidence of movement towards program outcomes, even if "soft" data need to be used.
5. Prepare the final report clearly and succinctly. The data and data interpretations should be presented in a manner that will help the reader recall the questions addressed and understand the nature and significance of the answers provided.

LEARNING EXERCISE 20: USE OF EVALUATION INFORMATION

There are a number of audiences for evaluation reports, some of which are listed below. Your group is to select an audience from the list or select another of your own choice, whichever you choose, your group should put itself in the position of that audience as you complete this exercise.

School Board
Community Group
Superintendent and Associates
Principal and Administration
Teacher Association
Parent Organization
Pupil Group
Other

As leaders in one or another of these groups, determine one or more purposes that you would want addressed in the program.

Read the final program evaluation report that begins on page H-5 and discuss it from your points of view. List as many actions, decisions, recommendations, or other uses that your group can act upon from the information supplied. List also some things that your group would like to have seen in the report but that were not included, and note the areas left without decision as a result of these shortcomings.

Make notes concerning your discussions on each of these three points on the exercise form on the following page.

EXERCISE FORM
USE OF EVALUATION INFORMATION

Audience

Purpose(s) for program evaluation:

Uses that could be made of the evaluation information, in priority:

Things that might have been included in the evaluation that would have been helpful to your audience. (Again, put in priority ranking):

Areas that are left without decisions as a result of these shortcomings:

SUNSET UNIFIED SCHOOL DISTRICT

DIAGNOSTIC AND PRESCRIPTIVE READING PROGRAM

JUNE 1976

PROGRAM EVALUATION REPORT

PROGRAM GOAL

The goal of the Diagnostic and Prescriptive Reading Program is to provide greater reading achievement gains for participating pupils than the traditional reading program provided for the same pupils during the previous year.

PROGRAM OBJECTIVES

1. Pupils participating in the Diagnostic and Prescriptive Reading Program will obtain an average gain of one month of reading achievement for each month of reading instruction as measured by pre-post testing with a standarized reading achievement test. The pupils' performance in the traditional reading program in 1974-75 yielded an average gain of one-half month of reading achievement per month of reading instruction.

PROGRAM DESCRIPTION

All pupils in grades one through six in Elmhurst, Diogenes and Mounthaven elementary schools in the Sunset Unified School District participated in the Diagnostic and Prescriptive Reading

Program during the school year of 1975-76.

Teachers utilized the Diagnostic/Prescriptive Teaching (DPT) approach to reading instruction as developed by the district's ESEA Title I Compensatory Education Program. This approach involves assessing each pupil to determine his current mastery level and the skills to be further mastered during the school year. The teacher then uses special materials designed to assist each pupil in his areas of need. After each unit of instruction, the teacher again assesses the pupil for mastery of the specific skills that were taught to determine whether the instruction was effective. If the pupil has mastered the skills in question, the teacher moves on to work with the pupil in his or her next area of need.

In this approach, teachers use a wide variety of instructional materials and equipment. Class size was limited to 28-30 pupils. Each teacher had an instructional aide for the purpose of assisting the pupils for three hours each day.

The program was in operation from November 1, 1975 to May 31, 1976 for a total of seven months of instructional time.

PROGRAM EVALUATION PROCEDURES

The Cooperative Primary Reading Test was administered to all first, second and third grade pupils by their classroom teachers on November 1, 1975 and again on May 31, 1976 for pre-post measurement of reading achievement. The Reading Test of the California Test of Basic Skills was administered to pupils in grades four, five and six by their teachers on the same dates as above.

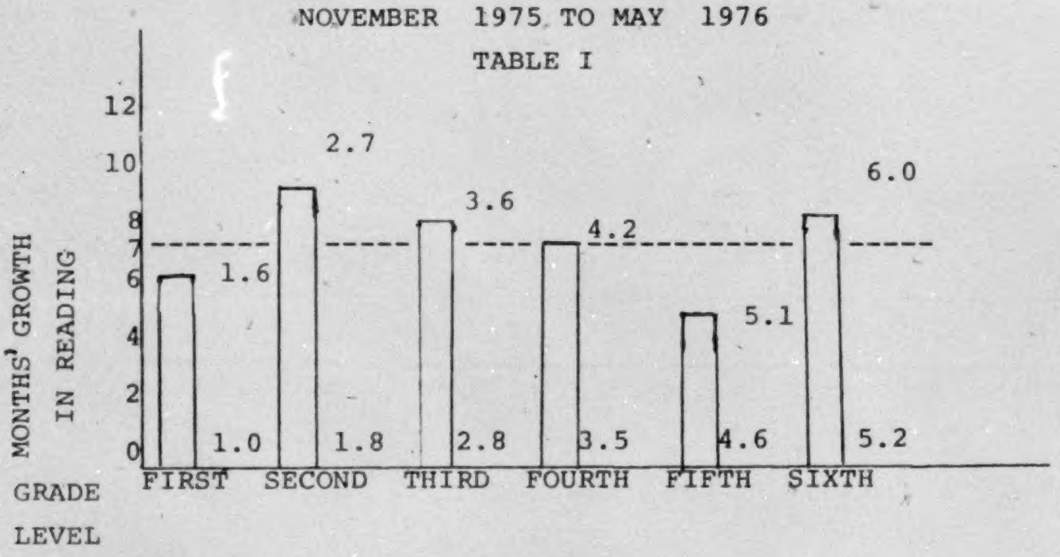
A questionnaire was developed and administered to each classroom teacher to survey their attitudes toward the program in May, 1976.

PROGRAM ACCOMPLISHMENTS

The program was implemented as described in the evaluation plan. The instructional aides were viewed as being helpful by the classroom teachers in each school. Learning prescriptions for pupils were developed for each pupil by the teacher after assessment of individual skill levels. With the exception of one school, learning centers for pupils were established and functioned as expected. Pupil testing was accomplished as scheduled and materials and equipment were provided as required in each school. Ongoing pupil records were adequately maintained as required for diagnostic prescriptive instruction.

The evaluation procedures determined to what extent the objective of an average gain of one month of reading achievement for each instructional month, or in other words, an average gain of seven months in reading achievement in a seven-months instructional period for each class in grades one through six, was accomplished. Table I gives the results in graphic form.

PUPILS' GAIN IN READING ACHIEVEMENT
NOVEMBER 1975 TO MAY 1976
TABLE I



It will be noted that in a seven-month period pupils in the first grade had a mean reading score of 1.0 in November and 1.6 in May. They, therefore, made an average gain of .6 or six months, which is one month short of the stated objective.

Second grade pupils begain the instructional program with an average reading score of 1.8 and ended with an average reading score of 2.7 with a mean gain of .9 or nine months, which is two months in excess of the stated objective of seven months.

Pupils in the third grade had a measured mean reading achievement score of 2.8 at the beginning of the program and 3.6 at the end with a mean gain of .8 or eight months which is one month in excess of the stated objective of seven months.

Fourth grade pupils had an average reading score of 3.5 at the beginning of the program and 4.2 at the end with a mean gain of .7 or seven months, which is identical to the stated objective of seven months.

Fifth grade pupils had an average reading score of 4.6 at the beginning of the program and 5.1 at the end with a mean gain of .5 or five months, which is two months less than the stated objective.

Pupils in the sixth grade began the instructional program with a mean reading score of 5.2 and ended with 6.0 with a mean gain of .8 or eight months, which is one month in excess of the stated objective of seven months.

In summary, pupils in grades two, three, four and six gained a mean score of seven months or more. Pupils in grades one and five did not meet the stated objective of seven months, though grade one missed by only one month and grade five by two months.

The overall mean gain of all pupils in grades one through six was 7.1 or slightly over seven months, which met the general objective of all pupils participating in the Diagnostic and Prescriptive Reading Program making an average gain of one month of reading achievement for each month of reading instruction as measured by pre-post testing with a standarized reading achievement test.

OTHER FINDINGS

A locally developed questionnaire administered to all participating teachers revealed that:

- Eighty percent of the teachers reported that more individualized attention could have been given to each pupil if the teachers had received more adult assistance in the classroom.
- Ninety percent of the teachers reported that their pupils had a wide range of academic weaknesses and that it was impossible to provide adequate assistance to each pupil.
- Sixty-five percent of the teachers requested inservice training in managing and grouping pupils for individualized instruction.

CONCLUSIONS

The Diagnostic/Prescriptive Teaching approach to reading instruction met the objectives as planned in grades one, two, three, four and six.

The first grade results suggest the possibility of problems in the instructional program at that level.

An analysis of the testing procedures at the fifth grade level revealed that different levels of the same test were used at pretest and post-test times. Use of inappropriate tests contaminated accurate reporting of pupil accomplishment at the fifth grade level and therefore interpretation of the data must be tentative.

RECOMMENDATIONS

1. The Diagnostic-Prescriptive Reading Program should be continued in grades one through six at Elmhurst Diogenes and Mounthaven elementary schools during the 1975-76 school years with appropriate attention to the stated recommendations.

2. The variability of achievement gains in the various grade levels should be further explored. Some grade levels seem to be benefiting more from the DPT approach than others. It would be well to consider a school-by-school analysis of the grade level data.

3. Explore the variability within schools in the reading achievement scores, particularly in those grades which did not meet the objective of seven months gain in a seven-months instructional program.

4. Investigate the situation of one school not providing learning cernters for pupils. Explore the possibility of testing learning centers vs. no learning centers in next year's evaluation design.

5. Alternate forms of the same level of the test should be used in pre and post-testing at all grade levels.

6. Consider the establishment of a cooperative teaching arrangement and allow pupils who lack certain skills to work with teachers who have special expertise in these areas.

7. Efforts should be made to increase the number of hours worked by instructional aides or to increase the number of aides.

8. Consider the use of volunteer parents as aides in the classroom.

9. Provide additional inservice education opportunities for teachers in the area of managing and grouping pupils for individualized instruction.

10. The evaluation of the achievement outcome of the Diagnostic Prescriptive Reading Program should be continued for the 1976-77 school year.

LEARNING EXERCISE 21: ROADBLOCKS TO PROGRAM EVALUATION

There are many reasons why the evaluations of educational programs are resisted. Participants will now be divided into "role" groups of three or four people each: board members, principals, classroom teachers, parents, superintendents, and so on.

Each group, looking at program evaluation from the view point of its role, should list as many roadblocks as possible to effective use of evaluation results. After these have been posted and reported on, the workshop leader will promote discussions of ways to overcome, circumvent, or minimize each roadblock identification.

PROGRAM EVALUATOR'S GUIDE

Section I

SELECTED BIBLIOGRAPHY

The Evaluation Improvement Program

SELECTED BIBLIOGRAPHY

Ahman, J. Stanley and Marvin D. Glock. *Evaluating Pupil Growth, Principles of Tests and Measurements*, 4th. Ed. Boston: Allyn and Bacon, Inc., 1971

Aiken, Jr., Lewis R. *Psychological and Educational Testing*, 1st Ed. Boston: Allyn and Bacon, Inc., 1971

Anderson, S. B., Ball, S., Murphy, R. T., and others. *Encyclopedia of Educational Evaluation*. San Francisco: Jossey-Bass, 1975.

Bloom, B. S. *Taxonomy of Educational Objectives: Cognitive Domain*. New York: David McKay, 1956.

Borich, G. D. (Ed.). *Evaluating Educational Programs and Products*. Englewood Cliffs, New Jersey: Educational Technology Publications, 1974.

Brown, Frederick G. *Measurement and Evaluation*, 1st Ed. Illinois: F. E. Peacock Publishers, Inc., 1971.

Brown, Frederick G. *Principles of Educational and Psychological Testing*. Illinois: The Dryden Press, Inc., 1970.

Cronbach, Lee J. *Essentials of Psychological Testing*, 2nd Ed. New York: Harper & Row, 1960.

Ebel, Robert L. *Measuring Educational Achievement*. Englewood Cliffs, New Jersey: Prentice Hall, 1965.

Feldt, L. S. "What Size Samples for Methods/Materials Experiments?", *Jr. of Ed. Measure*. Vol. 10, No. 3, 1973, 225.

Flanders, N. A. *Teacher Influence, Pupil Attitudes and Achievement*. Cooperative Research Monograph No. 12, OE 25040. Washington, D.C.: U.S. Government Printing Office, 1965.

Good, T. L. and J. E. Brophy. *Looking in Classrooms*. San Francisco: Harper & Row, 1973, pp. 62 and 63. Reprinted c 1975, Harper & Row Publishers.

Gronlund, Norman E. *Preparing Criterion-Referenced Tests for Classroom Instruction, A Title in the Current-Topics in Classroom Instruction Series*, New York: The MacMillian Company, 1973.

House, E. R. (Ed.). *School Evaluation: Politics and Process*. Berkeley: McCutchan, 1973.

Isaac, Stephen and William B. Michael. *Handbook in Research and Evaluation*. San Diego: Robert R. Knapp, 1971.

Krathwohl, D. R., B. S. Bloom and B. B. Masia. *Taxonomy of Educational Objectives*: Affective Domain. New York: David McKay, 1964.

Krejcie, Robert V. and Daryle W. Morgan. "Determining Sample Size for Research Activities," *Ed. & Psycho. Meas.*, Vol. 30, 1970, 607-610.

Kropp, R. P. and C. Verner. An attitude scale technique for evaluating meetings. *Adult Education*, VII(4), Summer, 1957.

Linton, M. and P.S. Gallo, Jr. *The Practical Statistican*: Simplified Handbook of Statistics. Monterey, California: Brooks Cole, 1975.

Loret, P.G., et al, Anchor Test Study: Equivalance and norms tables selected for reading achievement tests (grades 4,5, & 6). *Office of Education Report 74-305*, U.S. Government Printing Office, Washington, D.C., 1974.

Marshall, Jon Clark and Loyde Wesley Hales. *Essentials of Testing*, California" Addison-Wesley Publishing Company, 1972.

Marshall, Jon Clark and Loyde Wesley Hales. *Classroom Test Construction*, California: Addison-Wesley Publishing Company, 1971.

McFarland, Susan J. and Carl F. Hereford. *Statistics and Measurement in the Classroom, Psychological Foundations of Education Series*, 2nd Ed. Iowa: William C. Bloom Company Publishers, 1971.

Mehrens, William A. and Irvin J. Lehmann. *Measurement and Evaluation and Psychology*, California: Holt, Rinehart and Winston, Inc., 1973.

Metfessel, Newton S. and William B. Michael. *A Paradigm Involving Multiple Criterion Measures for the Evaluation of Effectiveness of School Programs, Educational and Psychological Measurement*. 1967.

National Assessment of Educational Progress. *Citizenship*: *National Results*. Denver, Colorado, November 1970.

Nunnally, Jr., Jum C. *Introduction to Psychological Measurement*, California: McGraw-Hill Book Company, 1970.

Payne, D. A. (Ed.). *Curriculum Evaluation*. Lexington, Massachusetts: D. C. Heath, 1974.

Popham, W. J. *Educational Evaluation*. Englewood Cliffs, New Jersey: Prentice Hall, 1975.

Provus, M. M. *Discrepancy Evaluation: For Educational Program Improvement and Assessment*. Berkeley: McCutchan, 1971.

Psychological Corporation. *Test Service Bulletin #48*. Printed by permission of the publisher, Psychological Corporation.

Simpson, E.J. *"The Classification of Educational Objectives, Psychomotor Domain"*. Illinois Teacher, 1966-67.

Simpson, E.J. *Scheme for Classification of Educational Objectives: Psychomotor Domain*. Unpublished Manuscript, 1969.

Slonin, M.J. Sampling: *A Quick, Reliable Guide to Practical Statistics*, New York: Simon & Schuster, 1967, 74.

Small-sample techniques. *The NEA Research Bulletin*, Vol. 38, December 1960, 99-104.

Smith, Fred M. and Sam Adams. *Educational Measurement for the Classroom Teacher*, 2nd Ed. California: Harper & Row, Publishers, 1972.

Stake, R. E. The Countenance of Educational Evaluation. *Teachers College Record*, 1967, *68*, 523-540.

Thorndike, Robert L. and Elizabeth Hagen. *Measurement and Evaluation in Psychology and Education*, 3rd Ed. New York: John Wiley and Sons, Inc., 1969.

Tyler, Leona E. *Tests and Measurements*. 2nd Ed. New Jersey: Prentice Hall, Inc., 1971.

Tyler, R. W., Gagne, R. M. & Scriven, M. Perspectives on Curriculum Evaluation. (American Education Research Association Monograph #1 in Curriculum Evaluation Series). Chicago: Rand-McNally, 1967.

Tyler, R. W. (Ed.). *Educational Evaluation: New Roles, New Means*. Chicago: University of Chicago Press, 1969.

Tyler, R. W. and Wolf, R. M. *Crucial Issues in Testing*. Berkeley: McCutchan, 1974.

Walberg, H. J. *Evaluating Educational Performance*. Berkeley: McCutchan, 1974.

Walker, Helen M. and J. Levy. *Statistical Inference*. New York: Holt, Rinehart and Winston, 1953.

Wick, John W. *Educational Measurement, Where Are We Going and How Will We Know When We Get There?* Ohio: Charles E. Merrill Publishing Company, 1973.

Womer, Frank B. *Guidance Monograph Series, Series III: Testing, Basic Concepts in Testing*, California: Houghton Mifflin Company, 1968.

Worthen, B. R. and Sanders, J. R. *Educational Evaluation: Theory and Practice*. Worthington, Ohio: C. A. Jones, 1973.

PROGRAM EVALUATOR'S GUIDE

Section J

APPENDICES

The Evaluation Improvement Program

CONTENTS

APPENDIX A: RESOURCES FOR INFORMATION ABOUT OBJECTIVES AND INSTRUMENTS . . . J-1

APPENDIX B: SELECTED LIST OF TEST PUBLISHERS J-10

APPENDIX C: MULTIPLE CRITERION MEASURES. J-12

APPENDIX A

RESOURCES FOR INFORMATION ABOUT OBJECTIVES AND INSTRUMENTS

I. Test Bulletins Published at Irregular Intervals

Normline, Harcourt Brace Jovanovich, Inc.
Test Data Reports, Harcourt Brace Jovanovich, Inc.
Test Service Bulletins, Harcourt Brace Jovanovich, Inc.
Test Service Bulletins, The Psychological Corporation
Test Service Notebook, Harcourt, Brace & World, Inc.
Testing Today, Houghton Mifflin Company

II. Newsletters

The ACT newsletter, American College Testing Program
Education Recaps, Educational Testing Service
ETS Developments, Educational Testing Service
Items, Cooperative Test Division, Educational Testing Service
Measurement in Education, National Council on Measurement in Education
NAEP Newsletter, National Assessment of Educational Progress
Test Collection Bulletin, Educational Testing Service

III. Educational and Psychological Journals

American Educational Research Journal
Education and Psychological Measurement
Journal of Educational Measurement
Journal of Educational Psychology
Measurement and Evaluation in Guidance
Psychological Abstracts: "Methodology and Research Technology: Testing" and "Educational Psychology: Testing"
Review of Educational Research

IV. Annual Reports and Proceedings

Annual Reports, College Entrance Examination Board
Annual Reports, Educational Testing Service
Proceedings, Annual Invitational Conference on Testing
Problems, Educational Testing Service
Proceedings, Annual Western Regional Conference on Testing
Problems, Educational Testing Service

V. Published Objectives and Objective-Referenced Tests

Instructional Objectives Exchange, Box 24095,
Los Angeles, California 90024
SCORE, Westinghouse Learning Corporation,
P.O. Box 30, Iowa City, Iowa 52240
National Evaluation Systems, P.O. Box 226,
Amherst, Massachusetts 01002

VI. Miscellaneous Paperback Books and Bulletins

EVALUATION AND ADVISORY SERIES, Educational Testing Service

1. ETS Builds A Test, 1965.
2. Locating Information on Educational Measurement: Sources and References, 1969.
3. Making the Classroom Test: A Guide for Teachers, Second Edition, 1961.
4. Multiple-Choice Questions: A Close Look, 1963.
5. Selecting an Achievement Test: Principles and Procedures, Second Edition, 1961.
6. Short-Cut Statistics for Teacher-Made Tests, Second Edition, 1964.

GUIDANCE MONOGRAPH SERIES, SET III: Testing, Houghton Mifflin Company, 1968

1. Modern Mental Measurement: A Historical Perspective
2. Basic Concepts in Testing
3. Types of Test Scores
4. School Testing Programs
5. Intelligence, Aptitude, and Achievement Testing
6. Interest and Personality Inventories
7. Tests on Trial
8. Automated Data Processing in Testing
9. Controversial Issues in Testing

Engelhart, M.D. Improving Classroom Testing, Washington: National Education Association, 1964

French, J.E., and W.B. Michael. Standards for Educational Psychological Tests and Manuals. Washington: American Psychological Association, 1966.

McLaughlin, K.F. Interpretation of Test Results. Washington: U.S. Government Printing Office, 1964.

VII. An Annotated Bibliography of Guides for Test Selection

Compiled by John Jegi, Director, ACCESS Information Center, Contra Costa County Superintendent of Schools Office

Buros, Oscar K., ed. Mental Measurements Yearbooks: 1st ed. - 1938 (re-issued 1972); 2nd ed. - 1941 (re-issued 1972); 3rd ed. - 1949; 4th ed. - 1953; 5th ed. - 1959; 6th ed. - 1965; 7th ed. - 1972 (2 vols.). Highland Park, New Jersey: The Gryphon Press.

Single best source of critical reviews of tests. Each yearbook contains critical reviews of all obtainable published tests and books on measurement written in English. Most publications are reviewed independently by two or more specialists. Reviews in earlier editions are cross-referenced in later ones.

______. Tests in Print. Highland Park, New Jersey: The Gryphon Press, 1961.

A comprehensive test bibliography and index to the first five books in the Mental Measurements Yearbook series. Each test mentioned includes information concerning test title, appropriate grade levels, publication data, special short comments about the test, number and types of scores provided, authors, publishers, and reference to test reviews in Mental Measurements Yearbooks.

______. Tests in Print. Volume 2. Highland Park, New Jersey: The Gryphon Press, 1974.

Index to tests still in print that are listed in all seven Mental Measurements Yearbooks. Includes bibliographies of references on the construction, use and validity of specific tests published through 1971.

______. Personality Tests and Reviews. Highland Park, New Jersey: The Gryphon Press, 1970.

Provides compilation of personality test reviews and specific test bibliographies listed in the first six Mental Measurements Yearbooks. Includes new material on personality testing, including a comprehensive bibliography of 513 personality tests and 7,116 new references dealing with the construction, use, and validity of specific tests. Also includes a master index to the nonpersonality tests, reviews and references to the first six Mental Measurements Yearbooks. Eighty tests--new, revised, or supplemented since the Sixth Yearbook and not listed in the Seventh Yearbook--are included.

______. Reading Tests and Reviews. Highland Park, New Jersey: The Gryphon Press, 1968.

Includes a comprehensive bibliography of reading tests as of early 1968, a reprinting of all reading test reviews in the first six Mental Measurements Yearbooks, and a master classified index to all other tests and reviews in the first six Yearbooks. Includes information about 33 reading tests--new, revised, or supplemented since the Sixth Yearbook which are not listed in the Seventh Yearbook.

Hoepfner, Ralph, Ed. CSE Elementary School Test Evaluations. Los Angeles: Center for the Study of Evaluation, UCLA, 1970.

This book contains a compendium of tests, keyed to educational objectives of elementary school education, and evaluated by measurement experts and educators for such characteristics as meaningfulness, examineee appropriateness, administrative usability, and quality of standardization.

Hoepfner, Ralph; Stern, Carolyn; and Nummendal, Susan G., eds. CSE-ECRC Preschool/Kindergarten Test Evaluations. Los Angeles: Center for the Study of Evaluation and the Early Childhood Research Center, UCLA, 1971.

This book contains a compendium of tests, keyed to educational objectives of early childhood education, and evaluated by measurement experts and educators.

Johnson, Orval G., and Bommarito, James. Tests and Measurements in Child Development: A Handbook. San Francisco: Jossey-Bass, Inc., Publishers, 1971.

A guide to more than 300 measures of child behavior and development not available from test publishers. Authors cite six criteria for inclusion: (1) suitability for use with children between birth and age twelve; (2) availability to professionals; (3) unpublished, not commercially available; (4) permit development of norms and reliability and validity data; (5) include enough information for effective use; (6) technically useable measures classified in ten categories: (a) cognitive, (b) personality and emotional characteristics, (c) children's perceptions of environment, (d) self-concept, (e) actual environment, (f) motor skills, brain injury, sensory perception, (g) physical attributes, (h) attitudes and interests not otherwise classified, (i) social behavior, and (j) measures not fitting the above categories.

Johnson, T.J., and Hess, R.J. Tests in the Arts. St. Louis: Central Midwestern Regional Educational Laboratory (CEMREL), 3120 - 59th Street, St. Louis, Mo., 1970.

Indexes and abstracts all known measuring instruments and tests applicable to the arts and provides a brief but comprehensive overview of the various psychometric methodologies utilized in the development of the instruments.

Robinson, John P., and Shaver, Philip R. Measures of Social Psychological Attitudes. Ann Arbor, Michigan: Publications Division, Institute for Social Research, University of Michigan, 1969*.

Review of 106 test instruments grouped into eight general categories: Life satisfaction; Self-esteem; Alienation; Authoritarianism; Socio-political attitudes; Values; General attitudes toward people; and Religious attitudes. Evaluation of instruments' psychometric properties given as well as ease of administration and scoring.

*Available in 1974 revised edition.

Simon, A. and E.G. Boyer. Mirrors for Behavior: An Anthology of Classroom Observations Instruments. Philadelphia: Research for Better Schools, 1967.

An annotated compilation of 86 observations instruments representating a variety of approaches, both in the affective and cognitive domains. Extensive bibliography.

Walker, Deborah K. Socioemotional Measures for Preschool and Kindergarten Children. San Francisco: Jossey-Bass, Inc. Publishers, 1973.

Description of 143 tests and measures of social and emotional development including titles and dates of publication or copyright; author; appropriate age range; measurement technique; source in which measure is described; description of the instrument; norms available; validity studies; and reliability evidence.

Wall, Janet, and Summerlin, Lee. Standardized Science Tests: A Descriptive Listing. Washington, D.C.: National Science Teachers' Association, 1973. (Order direct from NSTA, 1201 Sixteenth Street, NW, Washington, D.C. 20036 $1.50).

A compilation of virtually all the standardized science tests published since 1959 available to elementary and secondary science teachers. (57 pages).

The following documents are from the ERIC Clearinghouse on Tests, Measurement, and Evaluation, Educational Testing Service, Princeton, New Jersey and are available on microfiche. Items selected include clearinghouse publications through April, 1975.

ED 056 082. Rosen, Pamela, and Horne, Eleanor V. Language Development Tests: An Annotated Bibliography, 1971.

Brief annotations of currently available language development measures appropriate for use with preschool children as well as with lower elementary grade children (grades 1 through 3) are presented. The annotation provides information concerning the purpose of the test; the groups for which it is included; test subdivisions or tested skills, behaviors, or competencies; administration; scoring; interpretation; and standardization. (14 pages).

ED 056 083. Guthrie, P.D., and Horne, Eleanor V. School Readiness Measures: An Annotated Bibliography. 1971.

Brief Annotations of currently available general school readiness measures are presented. The annotation provides information concerning the purpose of the test; the groups for which it is intended; test subdivisions or tested skills, behaviors, or competencies; administration; scoring; interpretation; and standardization. An alphabetical listing of the instruments which indicates the ages for which each is suitable is also included. (26 pages).

ED 056 085. Guthrie, P.D., and others. Measures of Social Skills: An Annotated Bibliography, 1971.

Brief annotations of instruments concerned with a variety of social skills measures appropriate for use with children from the preschool level through the third grade are provided. Included are tests designed to measure social competency, interpersonal competency, social maturity, social sensitivity, and attitudes toward others. The annotation provides information concerning the purpose of the test; the groups for which it is intended; test subdivisions or tested skills; behaviors or competencies; administration; scoring; interpretation; and standardization. An age table is also provided which lists the tests alphabetically, indicates the ages for which each instrument is considered suitable, and gives the page on which each annotation appears. (28 pages).

ED 074 071. Knapp, Joan, Comp. An Omnibus of Measures Related to School-Based Attitudes. 1972.

Summaries are provided for 16 measures of school-based attitudes. All of the instruments are paper and pencil, self-report inventories. Some are designed for children 4-8 years of age; others are for students in grades 12-14. Each of the instruments is presented in the following format: Title, Description, Subjects, Response Mode, Scoring, and Comments. The 16 measures are: Survey of Study Habits and Attitudes; School Interest Inventory; The Student Opinion Poll II: School Morale Scale; Measures of School and Learning Attitudes; Attitudes Toward Education; Polittle Sentence Completion Test; Pictographic Self Rating Scale; Children's Attitudinal Range Indicator; When Do I Smile?; Attitude Toward Any School Subject; Attitude Instrument to Evaluate Student Attitude Toward Science and Scientists; Inventory of Reading Attitude; A Childhood Attitude Inventory for Problem Solving; Mathematics Attitude Scale; and a Semantic Differential for Measuring Attitudes of Elementary School Children Toward Mathematics. Fifteen references are provided. (24 pages).

ED 080 534. Knapp, Joan, Comp. A Selection of Self Concept Measures. 1973.

This compilation is comprised of descriptions of instruments for measuring self-concept. The instruments were chosen on the basis of the following criteria: they should be suitable for and reflect the full age range of children in school; each of the categories in Coller's model--self report, projective, behavior trace, and direct observation--should be represented; they should have been designed with the so-called "normal" population in mind rather than a psychopathological population; they have enough information accompanying them to enable investigators to use them effectively; and they should reflect a variety of means of presentation (e.g., pictorial items, semantic differential). The instruments described are: Work Posting; The Children's Self-Social Constructs Test; The Children's Self-Concept Index; Responsible Self-Concept Test; Behavior Rating Form; Coopersmith Self-Esteen Inventory; Tennessee Self-Concept Scale; How I See Myself Scale (Primary and Secondary Form); A Semantic Differential for Measurement of Global and Specific Self-Concepts; The Piers-Harris Children's Self-Concept Scale (The Way I Feel About Myself); Michigan State General Self-Concept of Ability, Michigan State Self-Concept of Ability in Specific Subjects Scales; and Self Esteem Measure for Neighborhood Youth Corps Enrolees. (31 pages).

ED 083 318. Rosen, Pamela, ed. Tests for Educationally Disadvantaged Adults. 1973.

Sixty-five instruments, published between 1925 and 1972, are described in this annotated bibliography. The devices are intended for adults who have received only an elementary education, and adults who have completed high school but whose education was impaired due to learning disabilities or other educational handicaps. Both achievement and aptitude measures are included, covering such areas as intelligence, ability, learning skills, non-verbal reasoning, vocabulary, reading, and mathematics. The Spanish editions of several tests in English as a second language are presented. The publisher's name and address is provided for each instrument. (12 pages).

ED 083 319. Rosen, Pamela, ed. Self-Concept Measures: Grade 7 and Above. 1973.

This 34-item annotated test bibliography deals with a variety of currently available measures of self-concept and self-esteem. For the purposes of this listing, self-concept was defined as a multi-dimensional construct encompassing the range of an individual's perceptions and evaluations of himself. Many of the devices contained herein emphasize the learner's self-concept or the individual's conceptions of himself in the school environment. However, several global measures are also described. Various methods for assessing self-concept, including direct observations, behavior ratings, self-reports, and projective techniques, are presented. The instruments described in this listing are appropriate for use in grade seven and above. Information was obtained from the holdings and references of the Educational Testing Service Test Collection. (7 pages).

ED 083 320. Rosen, Pamela, ed. *Measures of Self-Concept, Grades 4-6*. 1973.

This 31-item test bibliography deals with a variety of currently available measures of self-concept and self-esteem. For the purposes of this listing, self-concept was defined as a multi-dimensional construct encompassing the range of an individual's perceptions and evaluations of himself. Many of the devices contained herein emphasize the learner's self-concept of the child's conception of himself in the school environment. However, several global measures are also described. Various methods for assessing self-concept, including direct observations, behavior ratings, self-reports, and projective techniques, are presented. The instruments described in this listing are appropriate for use with children in grades four through six. Information was obtained from the holdings and references of the Educational Testing Service Test Collection. (6 pages).

ED 083 321. Rosen, Pamela, ed. *Attitudes Toward School and School Adjustment Grades 4-6*. 1973.

This 31-item test bibliography lists currently available measures of attitudes toward school and school adjustment. The construct--attitudes toward school--encompasses pupils' attitudes toward themselves as learners, learning as a process, the school environment or classroom situation, specific school subjects, and teachers. In addition, the pupils' behavior is considered if it is indicative of their adjustment or lack of adjustment to the educational environment. Teacher ratings, self-report devices, and observation techniques are the various methods for assessing these attitudinal elements which have been included in the listing. Instruments described in this bibliography are appropriate for use with students in grades four through six. Information was obtained from the holdings and references of the Educational Testing Service Test Collection. (8 pages).

ED 083 322. Rosen, Pamela, ed. *Assessment of Teachers*. 1973.

This 53-item test bibliography lists a variety of currently available measures which may be used to assess teachers. Among the devices described are: instruments which are completed by teachers and which provide an indication of their proficiency in or knowledge of both general and specific areas in education; self report attitudinal measures for teachers; instruments which are completed by students and which may indicate their attitudes toward and/or evaluations of a particular teacher or classroom istuation which is dependent upon the teacher; and observational devices that may be used to consider such factors as the teacher's competency, teaching style, characteristics and/or interaction with pupils. Information was obtained from the holdings and references of the Educational Testing Service Test Collection. (11 pages).

ED 083 323. Rosen, Pamela, ed. Attitudes Toward School and School Adjustment; Grades 7-12. 1973.

This 53-item test bibliography lists currently available measures of attitudes toward school and school adjustment. The construct--attitudes toward school--encompasses pupils' attitudes toward themselves as learners, learning as a process, the school environment or classroom situation, specific school subject, and teachers. In addition, the pupils' behavior is considered if it is indicative of their adjustment or lack of adjustment to the educational environment. Teacher ratings, self-report devices, and observational techniques are the various methods for assessing these attitudinal elements which have been included in the listing. Instruments described in this bibliography are appropriate for use with students in grades seven through twelve. Information was obtained from the holdings and references of the Educational Testing Service Test Collection. (7 pages).

ED 086 737. Rosen, Pamela. Self-Concept Measures. Head Start Test Collection. 1973.

Forty-four items published between 1963 and 1972 are listed in this annotated bibliography which deals with a variety of self-concept measures appropriate for use with children from the preschool level through the third grade. For the purposes of this listing, self-concept was defined as a multidimensional construct encompassing the range of a child's perceptions and evaluations of himself. Many of the sources emphasize the learner's self-concept or the child's conception of himself in the school environment. However, several global measures are also described. (8 pages).

ED 099 427. Knapp, Joan. A Collection of Criterion-Referenced Tests. TM Report No. 31. 1974.

Twenty-one criterion-referenced tests are cited and for each the following information is provided: description, format and administration, response mode and scoring, technical information, and references. The tests cited are the result of an attempt made to bring together tests designated in the Educational Testing Service Test Collection, a library of tests and test related information, and labeled in the ERIC system as criterion-referenced tests. This annotated bibliography does not list every test that has been labeled criterion-referenced; however, it typifies the variety of tests that are available under the rubric criterion-referenced. Also, criterion-referenced and norm-referenced tests are defined in several ways, and their advantages, limitations, and uses are beiefly explored. (13 pages).

APPENDIX B

SELECTED LIST OF TEST PUBLISHERS

AMERICAN COLLEGE TESTING PROGRAM, P.O. Box 168, Iowa City, Iowa 52240

AMERICAN GUIDANCE SERVICE, INC., Publishers' Building, Circle Pines,
Minnesota 55014

AUSTRALIAN COUNCIL FOR EDUCATIONAL RESEARCH, Frederick Street,
Hawthorn E.2, Victoria, Australia

BOBBS-MERRILL COMPANY, INC., 4300 West 62nd Street, Indianapolis,
Indiana 46268

BUREAU OF EDUCATIONAL RESEARCH AND SERVICE, University of Iowa,
Iowa City, Iowa 52240

CALIFORNIA TEST BUREAU/MCGRAW-HILL, Del Monte Research Park,
Monterey, California 93940

COMMITTEE ON DIAGNOSTIC READING TESTS, INC., Mountain Home,
North Carolina 28758

CONSULTING PSYCHOLOGISTS PRESS, INC., 577 College Avenue, Palo Alto,
California 94306

COOPERATIVE TESTS AND SERVICES, Educational Testing Service,
Princeton, New Jersey 08540

EDUCATIONAL AND INDUSTRIAL TESTING SERVICE, P.O. Box 7234, San Diego,
California 92107

EDUCATIONAL TEST BUREAU, Division of American Guidance Service,
Inc., 720 Eashington Avenue, S.E., Minneapolis, Minnesota 55414

EDUCATIONAL TESTING SERVICE, Princeton, New Jersey 08540

GUIDANCE CENTRE, Ontario College of Education, University of
Toronto, 1000 Yonge Street, Toronto 289, Ontario, Canada

HARCOURT BRACE JOVANOVICH, INC., 757 Third Avenue, New York,
New York 10017

HOUGHTON MIFFLIN COMPANY, 110 Tremont Street, Boston, Massachusetts
02107

INSTITUTE FOR PERSONALITY AND ABILITY TESTING, 1602 Coronado Drive,
Champaign, Illinois 61822

LYONS AND CARNAHAN, 407 East 25th Street, Chicago, Illinois 60616

PERSONNEL PRESS, INC., 20 Nassau Street, Princeton, New Jersey 08540

THE PSYCHOLOGICAL CORPORATION, 304 East 45th Street, New York,
New York 10017

PSYCHOMETRIC AFFILIATES, Box 31167, Munster, Indiana 46321

PUBLIC PERSONNEL ASSOCIATION, 1313 East 60th Street, Chicago,
Illinois 60637

SCHOLASTIC TESTING SERVICE, INC., 480 Meyer Road, Bensenville,
Illinois 60106

SCIENCE RESEARCH ASSOCIATES, INC., 259 East Erie Street, Chicago,
Illinois 60611

STANFORD UNIVERSITY PRESS, Stanford, California 94305

STOELTING COMPANY, 424 North Homan Avenue, Chicago, Illinois 60624

TEACHERS COLLEGE PRESS, Teachers College, Columbia University,
New York, New York 10027

APPENDIX C

MULTIPLE CRITERION MEASURES[1]

A. Indicators of Status or Change in Cognitive and Affective Behaviors of Students in Terms of Standardized Measures and Scales.

1. Standardized achievement and ability tests, the scores on which allow inferences to be made regarding the extent to which cognitive objectives concerned with knowledge, comprehension, understanding, skills and applications have been attained.

2. Standardized self-inventories designed to yield measures of adjustment, appreciations, attitudes, interests, and temperament from which inferences can be formulated concerning the possession of psychological traits (such as defensiveness, rigidity, aggressiveness, cooperativeness, hostility, and anxiety).

3. Standardized rating scales and checklists for judging the quality of products in visual arts, crafts, shop activities, penmanship, letter-writing, fashion design, and other activities.

B. Indicators of Status or Change in Cognitive and Affective Behaviors of Students by Informal or Semiformal Teacher-made Instruments or Devices.

1. Interviews: frequencies and measurable levels of responses to formal and informal questions raised in a face-to-face interrogation.

2. Questionnaires: frequencies of responses to items in an objective format and numbers of responses to categorized dimensions developed from the content analysis of responses to open-ended questions.

3. Self-concept perceptions: measures of current status and indices of congruence between real self and ideal self often determined from use of the semantic differential or Q-sort techniques.

---

[1] Metfessel, Newton S. & Michael, William B "A Paradigm Involving Multiple Criterion Measures for the Evaluation of Effectiveness of School Programs", "Educational & Psychological Measurement", 1967, p. 27, 931-943.

4. Self-evaluation measures: student's own reports on his perceived or desired level of achievement, on his perceptions of his personal and social adjustment, and on his future academic and vocational plans.

5. Teacher-devised projective devices such as casting characters in the class play, role playing, and picture interpretation based on an informal scoring model that usually embodies the determination of frequencies or the occurrence of specific behaviors, or ratings of their intensity or quality.

6. Teacher-made achievement tests (objective and essay), the scores on which allow inferences regarding the extent to which specific instructional objectives have been attained.

7. Teacher-made rating scales and check lists for observation of classroom behaviors; performance levels of speech, music and art; manifestation of creative endeavors, personal and social adjustment, physical wellbeing.

8. Teacher-modified forms (preferably with consultant aid) of the semantic differential scale.

C. Indicators of Status or Change in Student Behavior Other Than Those Measured by Tests, Inventories, and Observation Scales in Relation to the Task of Evaluating Objectives of School Programs.

1. Absences: full-day, half-day, part-day, and other selective indices pertaining to frequency and duration of lack of attendance.

2. Anecdotal records: critical incidents noted including frequencies of behaviors judged to be highly undesirable or highly deserving of commendation.

3. Appointments: frequencies with which they are kept or broken.

4. Articles and stories: numbers and types published in school newspapers, magazines, journals, or proceedings of student organizations.

5. Assignments: numbers and types completed with some sort or quality rating or mark attached.

6. Attendance: frequency and duration when attendance is required or considered optional (as in club meetings, special events, or off-campus activities).

7. Autobiographical data: behaviors reported that could be classified and subsequently assigned judgmental values concerning their appropriateness relative to specific objectives concerned with human development.

8. Awards, citations, honors, and related indicators of distinctive or creative performance: frequency of occurrence or judgments of merit in terms of sealed values.

9. Books: numbers checked out of library, numbers renewed, numbers reported read when reading is required or when voluntary.

10. Case histories: critical incidents and other passages reflecting quantifiable categories of behavior.

11. Changes in program or in teacher as requested by student: frequency or occurrence.

12. Choices expressed or carried out: vocational, avocational, and educational (especially in relation to their judged appropriateness to known physical, intellectual, emotional, social, aesthetic, interest, and other factors.)

13. Citations: commendatory in both formal and informal media of communication such as in the newspaper, television, school assembly, classroom, bulletin board, or elsewhere (see Awards).

14. "Contract": frequency or duration of direct or indirect communications between persons observed and one or more significant others with specific reference to increase or decrease in frequency or to duration relative to selected time intervals.

15. Disciplinary actions taken: frequency and type.

16. Dropouts: numbers of students leaving school before completion of program of studies.

17. Elected positions: numbers and types held in class, student body, or out-of-school social groups.

18. Extracurricular activities: frequency or duration of participation in observable behaviors amenable to classification such as taking part in athletic events, charity drives, cultural activities, and numerous service-related avocational endeavors.

19. Grade placement: the success or lack of success in being promoted or retained; number of times accelerated or skipped.

20. Grade point average: including numbers of recommended units of course work in academic as well as in non-college preparatory programs.

21. Grouping: frequency and/or duration of moves from one instructional group to another within a given class grade.

22. Homework assignments: punctuality of completion, quantifiable judgments of quality such as class marks.

23. Leisure activities: numbers and types of; times spent in; awards and prizes received in participation.

24. Library card: possessed or not possessed; renewed or not renewed.

25. Load: numbers of units or courses carried by students.

26. Peer group participation: frequency and duration of activity in what are judged to be socially acceptable and socially undesirable behaviors.

27. Performance: awards, citations received; extra credit assignments and associated points earned; numbers of books or other learning materials taken out of the library, products exhibited at competitive events.

28. Recommendations: numbers of and judged levels of favorableness.

29. Recidivism by students: incidents (presence or absence or frequency of occurrence) of a given student's returning to a probationary status, to a detention facility, or to observable behavior patterns judged to be socially undesirable (intoxicated state, dope addiction, hostile acts including arrests, sexual deviation).

30. Referrals: by teacher to counselor, psychologist, or administrator for disciplinary action, for special aid in overcoming learning difficulties, for behavior disorders, for health defects or, for part-time employment activities.

31. Referrals: by student himself (presence, absence, or frequency).

32. Service points: numbers earned.

33. Skills: demonstration of new of increased competencies such as those found in physical education, crafts, homemaking, and the arts that are not measured in a highly-valid fashion by available tests and scales.

34. Social mobility: numbers of times student has moved from one neighborhood to another and/or frequency with which parents have changed jobs.

35. Tape recordings: critical incidents contained and other analyzable events amenable to classification and enumeration.

36. Tardiness: frequency of.

37. Transiency: incidents of.

38. Transfers: numbers of students entering school from another school (horizontal move).

39. Withdrawal: numbers of students withdrawing from school or from a special program (see Dropouts).

D. Indicators of Status or Change in Cognitive and Affective Behaviors of Teachers and Other School Personnel in Relation to the Evaluation of School Programs.

1. Articles: frequency and types of articles and written documents prepared by teachers for publication or distribution.

2. Attendance: frequency of, at professional meetings or at inservice training programs, institutes, summer schools, colleges and universities (for advanced training) from which inferences can be drawn regarding the professional person's desire to improve his competence.

3. Elective offices: numbers and types of appointments held in professional and social organizations.

4. Grade point average: earned in postgraduate courses.

5. Load carried by teacher: teacher-pupil or counselor-pupil ration.

6. Mail: frequency of positive and negative statements in written correspondence about teachers, counselors, administrators, and other personnel.

7. Memberships including elective positions held in professional and community organizations: frequency and duration of association.

8. Model congruence index: determination of how well the actions of professional personnel in a program approximate certain operationally-stated judgmental criteria concerning the qualities of a meritorious program.

9. Moonlighting; frequency of outside jobs and time spent in these activities by teachers or other school personnel.

10. Nominations by peers, students, administrators or parents for outstanding service and/or professional competencies: frequency of.

11. Rating scales and checklists (e.g., graphic rating scales or the semantic differential) of operationally-stated dimensions of teachers' behaviors in the school setting from which observers may formulate inferences regarding changes of behavior that reflect what are judged to be desirable gains in professional competence, skills, attitudes, adjustment, interests, and work efficiency; the perceptions of various members of the total school community (parents, teachers, administrators, counselors, students, and classified employees) of the behaviors of other members may also be obtained and compared.

12. Records and reporting procedures practiced by administrators, counselors, and teachers: judgments of adequacy by outside consultants.

13. Termination: frequency of voluntary or involuntary resignation or dismissals of school personnel.

41. Transfers: frequency of requests of teachers to move from one school to another.

E. Indicators of Community Behaviors in Relation to the Evaluation of School Programs.

1. Alumni participation: numbers of visitations, extent of involvement in PTA activities, amount of support of a tangible (financial) or a service nature to a continuing school program or activity.

2. Attendance at special school events, at meeting of the board of education, or at other group activities by parents: frequency of.

3. Conference of parent-teacher, parent-counselor, parent-administrator sought by parents: frequency or request.

4. Conferences of the same type sought and initiated by school personnel: frequency of requests and record of appointments kept by parents.

5. Interview responses amenable to classification and quantification.

6. Letters (mail): frequency of requests for information, materials, and servicing.

7. Letters: frequency of praiseworthy or critical comments about school programs and services and about personnel participating in them.

8. Participant analysis of alumni: determination of locale of graduates, occupation, affiliation with particular institutions, or outside agencies.

9. Parental response to letters and report cards upon written or oral request by school personnel: frequency of compliance by parents.

10. Telephone calls from parents, alumni, and from personnel in communications media (e.g., newspaper reports): frequency, duration, and quantifiable judgments about statements monitored from telephone conversations.

11. Transportation requests: frequency of.